Three factors in explaining scalar diversity

Eszter Ronai & Ming Xiang

The University of Chicago

Sinn und Bedeutung (SuB) 26
8-10 September 2021

Introduction

In conversation, comprehenders draw inferences beyond literal meaning: **scalar inference**, e.g. *some but not all*.

Scalar diversity: **likelihood** of drawing such an inference **varies across scales**.

We explore factors that may explain this variation, and find that the following all contribute:

- How accessible an alternative like *all* is.
- How distinct *some* and *all* are.
- What the negated *not all* means.

Roadmap

1. Background
   1. Scalar inference.
   2. Scalar diversity.
2. Replication of scalar diversity.
3. Experiment 1: Accessibility of the stronger alternative.
4. Experiment 2: Distinctness of scalemates.
5. Experiment 3: The meaning of the negated strong scalar.
6. Conclusions.

Scalar inference

Scalar inference (SI) calculation:

(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.

Comprehenders reason about what is not said: the stronger alternative

- *all* in (1)
- *brilliant* in (2)

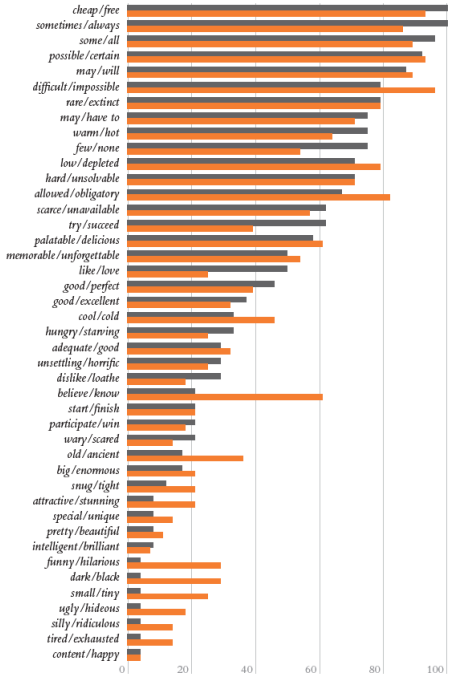
(Grice 1967)

Scalar diversity

Considerable **variation across different scales in SI calculation rates**.

E.g. *some but not all* SI arises much more robustly than *intelligent but not brilliant* —finding about 43 scales (van Tiel et al. 2016; see also Doran et al. 2012; Beltrama & Xiang 2013).

43 scales tested by van Tiel et al.



Explaining scalar diversity

What properties of scales can explain this variation?

- Semantic distance between scalars (van Tiel et al. 2016).
- Boundedness of the scale (van Tiel et al. 2016).
- Local enrichability (Sun et al. 2018).
- Extremeness (Gotzner et al. 2018; Beltrama & Xiang 2013).
- Polarity (Gotzner et al. 2018).
- Negative strengthening (Gotzner et al. 2018).
- Availability of the relevant QUD (Ronai & Xiang, 2021).
- The relevance of the SI (Pankratz & van Tiel 2021).

But: a lot of the (statistical) variance is still unaccounted for.
That is, **a lot of scalar diversity is unexplained**.

Research goals

Three factors investigated:

- **Accessibility of the stronger alternative**, given the weaker scalar.
  - Measured via a cloze production task.
- **Distinctness of** the two **scalar terms**.
  - Measured via degree estimates: “weak” vs. “strong”.

→ Inherent **properties of the relation between the weak and the strong scalar**.

- Meaning of the **the negated strong scalar** term, as compared to the weak scalar.
  - Measured via degree estimates: “weak” vs. **“not strong”**.

Collecting lexical scales: corpus study

Previous work: mostly (70%, e.g. van Tiel et al.) or entirely (e.g. Gotzner et al.; Pankratz & van Tiel) on adjectival scales. → Our aim: **better balance** across grammatical categories.

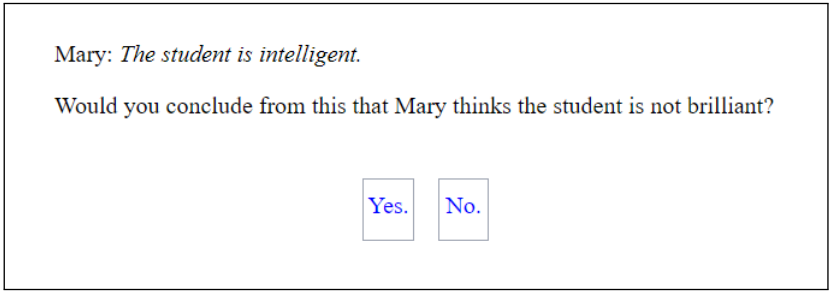
Scale sets from Marneffe & Tonhauser 2019 and van Tiel et al. 2016
+ COCA searches: *X or even Y*; *not just X but Y*; *X but not Y* (for adjectives, verbs, adverbs).

Filter: semantic tests for asymmetric entailment and cancellability.

Final set: 60 lexical scales.

Replication of scalar diversity

- 40 native speakers of American English; MTurk; IbexFarm.
- **Inference task**: test the likelihood of SI derivation from the 60 scales.

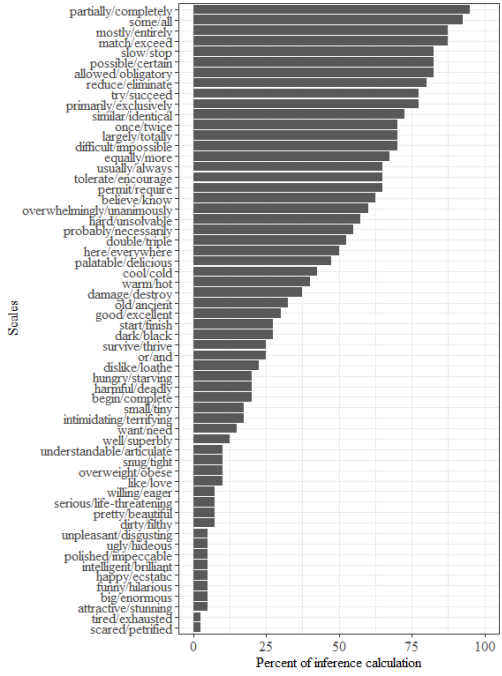


- "Yes" response = SI was calculated; "No" response = SI was not calculated.

Replication of van Tiel et al. (2016)

Results

- Scalar diversity replicated.



Experiment 1: accessibility of stronger alternative

Hypothesis: scalar diversity can (in part) be explained by **how accessible a stronger alternative is**, given the weaker scalar.

Causal mechanism behind our hypothesis:

- SI proceeds via **reasoning about alternatives**.
- Hearers **generate** a set of **alternatives**.
- The more accessible the alternative, the more likely hearers are to reason about it, and therefore the more likely the SI.

Experiment 1: accessibility of stronger alternative

Intuition: there may be differences across scales in how strongly the weaker scalar evokes a stronger alternative.

- *some*: *all* always comes to mind
- *intelligent*: a number of competing alternatives may be activated, such as *brilliant, hardworking, kind, crafty*, etc.

Relation to alternative availability

Van Tiel et al.'s hypothesis: the **availability** of the stronger alternative should predict scalar diversity.

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 she uttered.

Previous work testing alternative availability

Van Tiel et al.'s operationalization of availability:

- Association strength between weaker and stronger scalar (production-based).
- Grammatical class (open vs. closed).
- Frequency (relative, and absolute of stronger scalar).
- Semantic relatedness between weaker and stronger scalar (LSA score).

None of the above found to be a predictor of diversity.

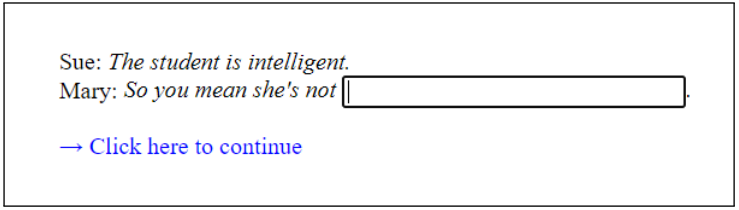
Experiment 1: metric for accessibility

Our operationalization: **cloze probability**, commonly used to measure the predictions the parser makes in language comprehension.

Probability of a target word completing a particular sentence frame, indexing how expected a word is in a context (Taylor, 1953; see also i.a. Kutas & Hillyard, 1984)

Experiment 1: accessibility of stronger alternative

- Modified cloze task: participants instructed to complete the answer with the first word that comes to mind.

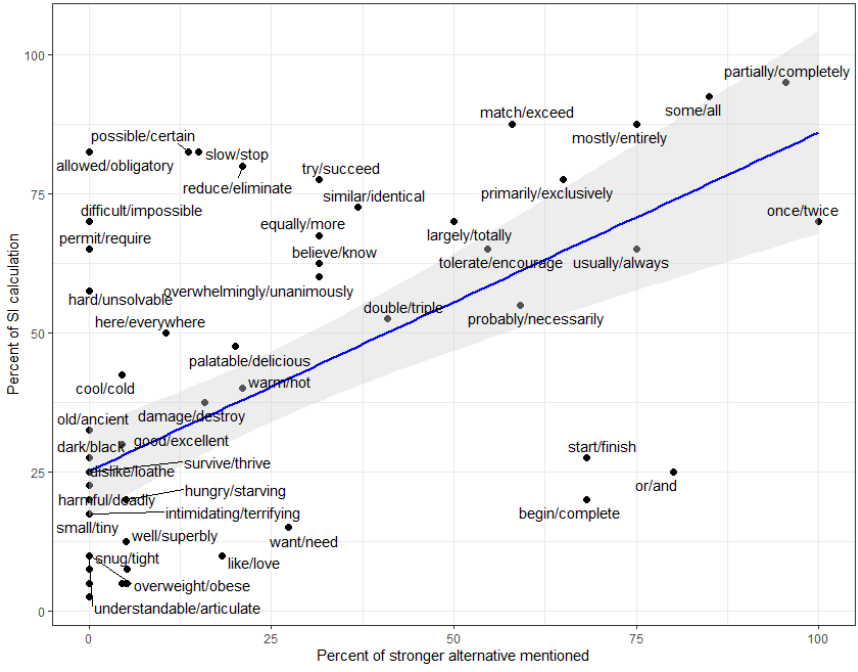


- 61 native speakers of American English → 19-22 completions per scale.
- MTurk, IbexFarm.
- **Prediction:** the **more frequently the stronger alternative is mentioned** in the cloze task, the **higher the SI rate** for that scale.

Results

% of mentioning the stronger alternative predicts SI rate ($p < 0.001$)

(coding of results counted synonyms)



Experiment 1: accessibility of stronger alternative

Scalar diversity: **predicted by the accessibility** of the stronger scalar, i.e. how strongly a weaker scalar evokes a stronger alternative.

Potential caveat: our measure of accessibility may be interpreted as the production-side of scalar diversity. Outcomes of the same mechanism?

Experiment 2: distinctness of scalar terms

Distinctness of two scalar terms as a predictor of scalar diversity (van Tiel et al.).

- SI is the negation of the stronger alternative (*not all, not brilliant*).
- The speaker could have uttered a stronger alternative, but she didn't, so it's not true.
- For this reasoning to go through, there has to be a **clear stronger alternative**, and it has to be **sufficiently stronger**.
- If it's **difficult to distinguish the weak and strong** scalar ("near-synonyms"), **SI is unlikely**.

Experiment 2: metric for distinctness

Operationalization:
inspired by Bayesian pragmatics, which assumes and models recursive reasoning between speaker and hearer (Goodman & Frank 2016; Lassiter & Goodman 2015; Xiang et al. under review).

→ Collect empirical data on **what information hearers think is communicated** by utterances that contain scalar terms.

Experiment 2: metric for distinctness

Speaker: *The student is intelligent/brilliant.*
What world states do hearers think such utterances describe?

→ Experimentally collect **degree estimates on the underlying degree scales**. To what degree do hearers think the student is intelligent?

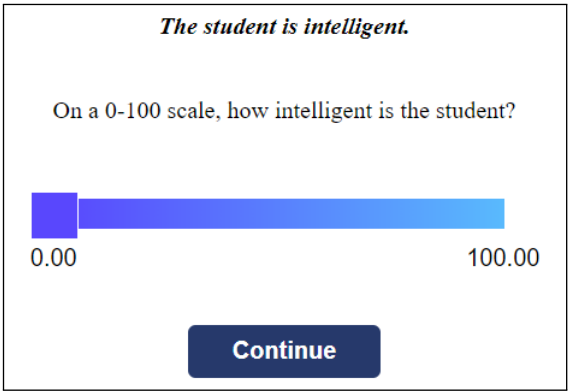
Experiment 2: distinctness of scalar terms

Hypothesis: the **bigger the difference between the weak and the strong** term, i.e. the further apart they are on the underlying degree scale, the **higher the SI rate** for that scale.

For an SI (*intelligent* → *not brilliant*) to arise, *intelligent* and *brilliant* have to be perceived as describing two different world states.

Experiment 2: distinctness of scalar terms

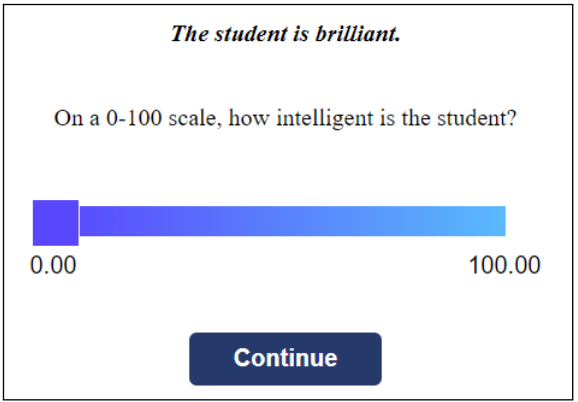
- 30 native speakers of American English; MTurk; IbexFarm.
- Degree estimate task: participants instructed to answer the question by picking a point on a scale from 0 to 100. —judgment on **weaker** scalar term



Data: represents **hearers' probabilistic guesses** on what **world state the speaker has in mind**, given her utterance.

Experiment 2: distinctness of scalar terms

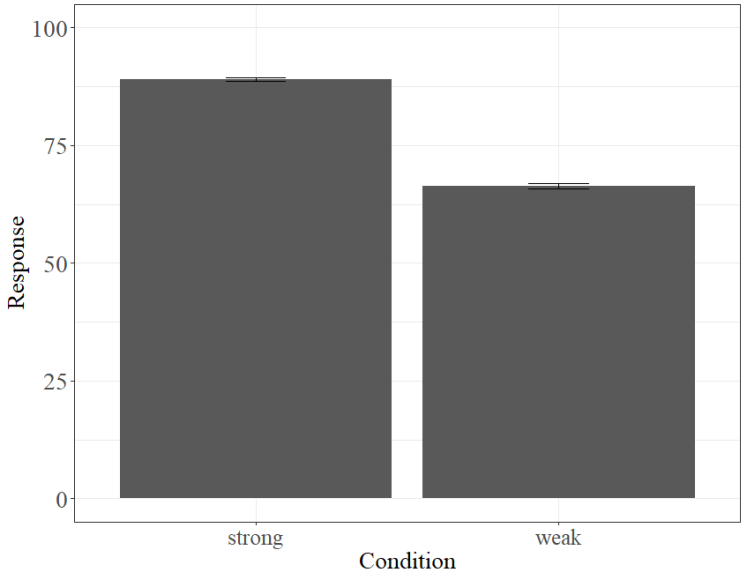
- 30 native speakers of American English; MTurk; IbexFarm.
- Degree estimate task: participants instructed to answer the question by picking a point on a scale from 0 to 100. —judgment on **stronger** scalar term



A note on degree estimates

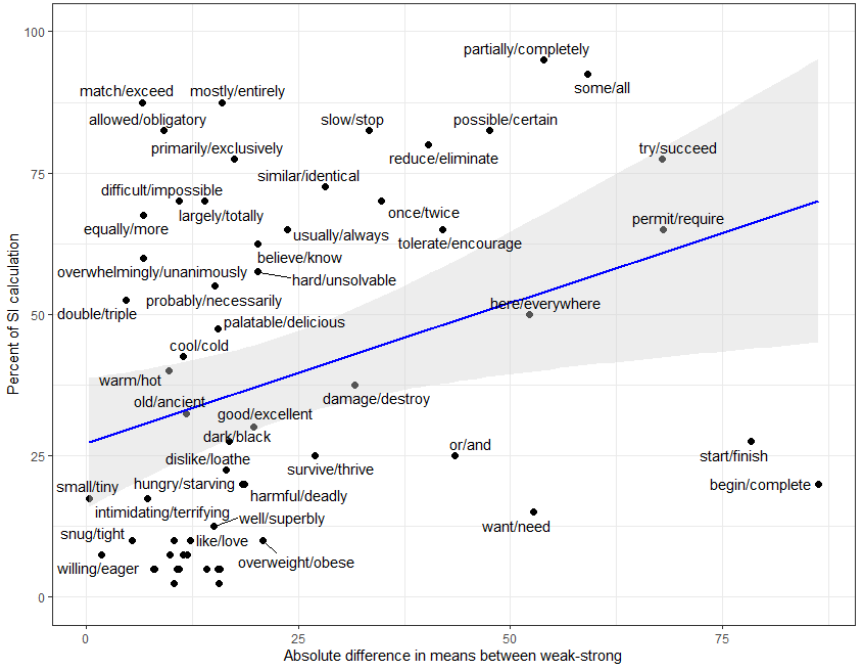
Caveat: our task is an idealization in that not all lexical scales map onto a bounded underlying degree scale.

Results:
2 conditions



Significant effect of condition ($p < 0.001$).

Results



weak-strong difference predicts SI rate ($p < 0.001$)

Experiment 2: distinctness of scalar terms

Difference between **weak** and **strong** term: positively correlated with SI rate.

→ The **more distinct the world states** that the weaker and the stronger term are taken to describe, the **higher the SI rate** for that scale.

Relation to semantic distance

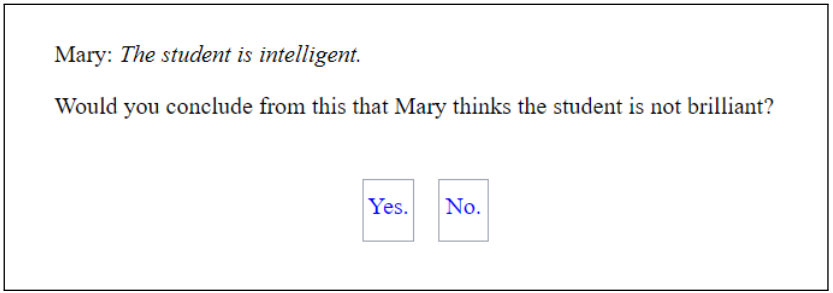
Semantic distance: the more distant a weak and a strong scalar term, the more likely the SI.

Part of van Tiel et al.'s operationalization of distinctness —happy to discuss more in the Q&A.

Shifting gears

So far: two factors investigating the relationship between the weak and the strong scalar term.

Let's consider the SI calculation inference task:



Probe: what does "*not brilliant*" even mean? What do people have in mind when answering this question?

Experiment 3: meaning of the negated strong scalar

Core idea: the **meaning of the negated stronger predicate** (e.g. *The student is not brilliant*) also matters for scalar diversity.

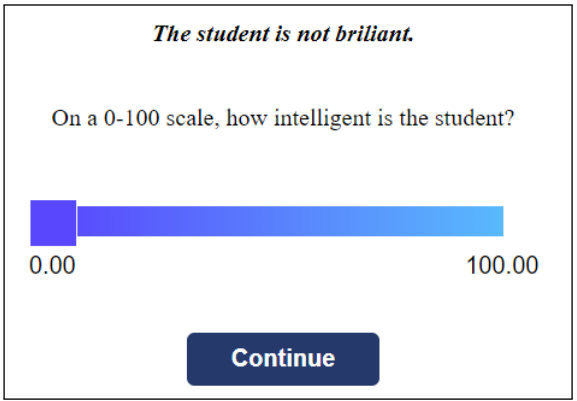
Hypothesis: the **smaller the difference between the weak and the negated strong** term, i.e. the closer they are on the degree scale, the **higher the SI rate** for that scale.

If *intelligent* and *not brilliant* are interpreted as describing two very different world states → it is implausible to conclude that the speaker meant *not brilliant* when she uttered *intelligent*.

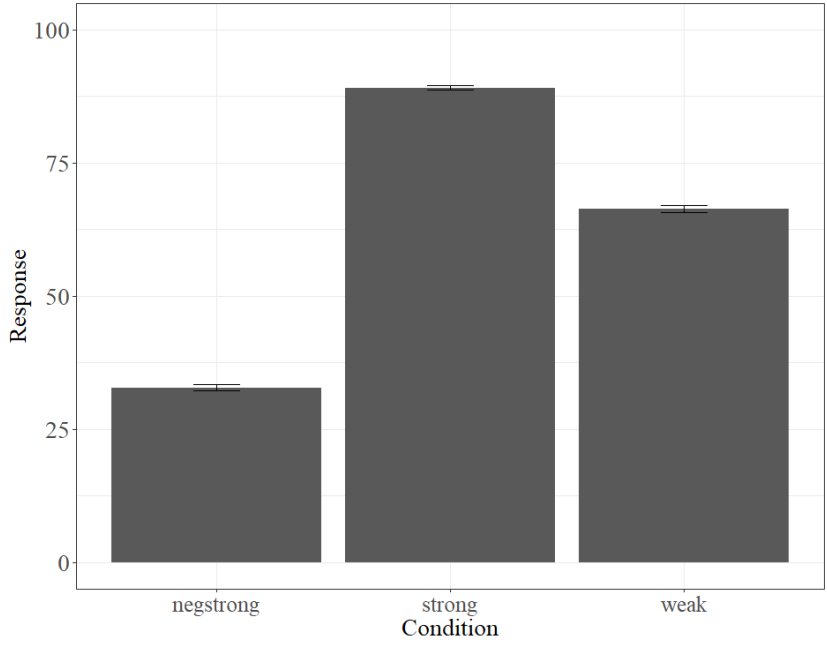
Metric: degree estimates.

Experiment 3: meaning of the negated strong scalar

- 31 native speakers of American English; MTurk; IbexFarm.
- Degree estimate task: participants instructed to answer the question by picking a point on a scale from 0 to 100. —judgment on **negated stronger** scalar term

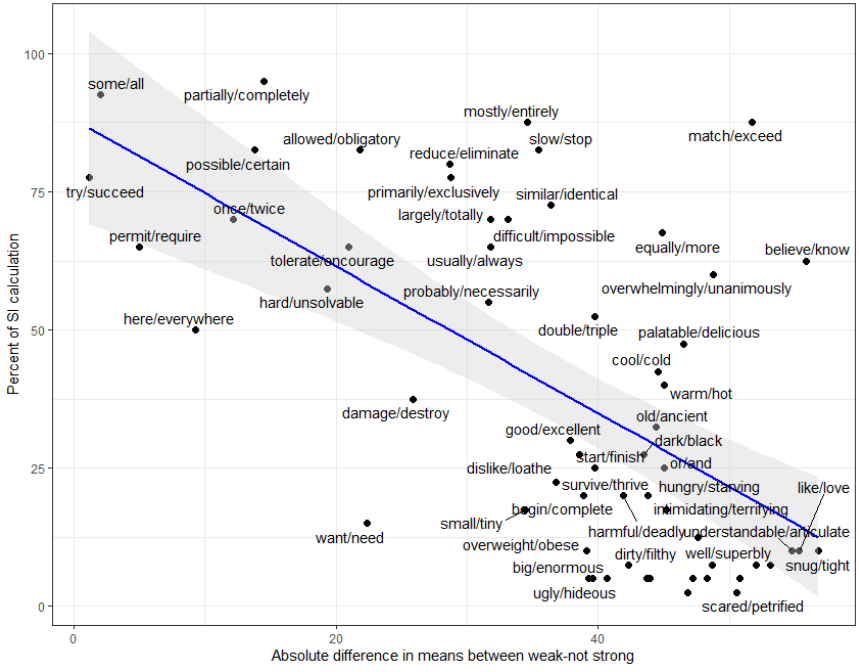


Results: 3 conditions



Significant effect of condition (all $p < 0.001$).

Results



weak-not strong difference predicts SI rate ($p < 0.001$)

Experiment 3 results: meaning of the negated strong scalar

Difference between **weak** and **negated strong** term: negatively correlated with SI rate.

→ The **more similar the world states** that the weaker and the negated stronger term are taken to describe, the **higher the SI rate** for that scale.

Relation to negative strengthening

For many scales, the "negated strong" degree estimate was lower than the "weak" degree estimate → may remind you of **negative strengthening** (Horn, 1989):

e.g. *John is not stunning* is interpreted as conveying that John is rather ugly.

Experimentally tested by Gotzner et al. (2018) —happy to discuss more in the Q&A.

Taking stock

Looking at the properties of the relation between the weak and strong scalar:

- The **accessibility** of the stronger alternative matters (as measured via a cloze task).
- The **distinctness** of the weak and strong scalar terms matters (as measured via degree estimates).

The **meaning** of the **negated stronger alternative**, vis-à-vis the weak scalar, also matters (as measured via degree estimates).

How much of the variance is explained?

Model with all three predictors: R^2 =22.4% (fixed effects only)

Variance accounted for by each factor:

- Accessibility of stronger scalar: R^2 =7.9%
- Distinctness between weak-strong: R^2 =2.7%
- Meaning of the negated strong scalar: R^2 =9.4%

(Test: how much is the R^2 reduced by taking that predictor out of the model?)

Future work: synthesis of all predictors of scalar diversity and total variance accounted for.

Conclusions

- We replicate scalar diversity on 60 scales that span grammatical categories.
- Three factors to capture scalar diversity:
  - Alternative accessibility, via a cloze task.
  - Distinctness of alternatives, via degree estimates.
  - Meaning of the negated strong scalar (vis-à-vis the weak scalar), via degree estimates.

Thank you!

ronai@uchicago.edu
mxiang@uchicago.edu

References

Beltrama & Xiang (2013). Is 'good' better than 'excellent'? An experimental investigation on scalar implicatures and gradable adjectives. Proceedings of SuB 17.

de Marneffe & Tonhauser (2019). Inferring meaning from indirect answers to polar questions. Questions in Discourse.

Doran et al. (2012). A novel experimental paradigm for distinguishing between what is said and what is implicated. Language.

Goodman & Frank (2016). Pragmatic language interpretation as probabilistic inference. Trends in Cognitive Science.

Gotzner et al. (2018). Scalar Diversity, Negative Strengthening, and Adjectival Semantics. Frontiers in Psychology.

Grice (1967). Logic and conversation. Studies in the way of words.

Horn (1972). On the Semantic Properties of Logical Operators in English. Ph.D. thesis.

Horn (1989). A Natural History of Negation. University of Chicago Press.

Kutas & Hillyard (1984). Brain potentials during reading reflect word expectancy and semantic association. Nature.

Lassiter & Goodman (2015). Adjectival vagueness in a Bayesian model of interpretation. Synthese.

Pankratz & van Tiel (2021). The role of relevance for scalar diversity: a usage-based approach. Language and Cognition.

Ronai & Xiang (2021). Exploring the connection between Question Under Discussion and scalar diversity. Proceedings of the LSA.

Sun et al. (2018). A link between local enrichment and scalar diversity. Frontiers in Psychology.

Taylor (1953). "Cloze procedure": a new tool for measuring readability. Journalism Quarterly.

van Tiel et al. (2016). Scalar diversity. Journal of Semantics.

Xiang et al. (under review). Pragmatic reasoning and semantic convention: a case study of gradable adjectives.

Relation to semantic distance

Semantic distance: the more distant a weak and a strong scalar term, the more likely the SI. Part of van Tiel et al.'s operationalization of distinctness.

(3) a. Many of the senators voted against the bill.
b. Most of the senators voted against the bill.
c. All of the senators voted against the bill.

SI from (3a): more likely the negation of (3c) than of (3b) (Horn, 1972).

Relation to semantic distance

Van Tiel et al.: participants rated how much stronger (1=equally strong to 7=much stronger) *She is brilliant* is than *She is intelligent* —positively correlated with SI rates.

Differences:

- Our experiments don't a priori assume a strength relation.
- Not relying on metalinguistic judgments yields a more natural task.

Relation to negative strengthening

Negative strengthening (Horn, 1989):
e.g. *John is not stunning* is interpreted as conveying that John is rather ugly.

Experimentally tested by Gotzner et al. (2018):
Participants saw *He is not brilliant* + asked whether they can conclude "He is not intelligent".
"Yes" responses negatively correlated with SI rates.

Differences:

- Our results **include scales that did not show negative strengthening**: negated strong scalar had a higher mean on the 0-100 scale than the weak scalar.
- Negative strengthening is chiefly about *not brilliant* being lower on the intelligence scale than *intelligent*.
- What our measure is about is *not brilliant* being similar to *intelligent*.