Quantifying scalar diversity: a first look

Eszter Ronai & Ming Xiang

The University of Chicago

Architectures and Mechanisms for Language Processing (AMLaP) 2021
2-4 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**.

This talk:

- First step towards quantifying this variation (relative entropy).
- What (semantic vs. pragmatic) manipulations can reduce this variation.

Roadmap

1. Background
   1. Scalar inference.
   2. Scalar diversity.
2. Experiment 1: Replication of scalar diversity + semantic manipulation (*only*).
3. Experiment 2: Pragmatic manipulation (Question Under Discussion).
4. Quantifying scalar diversity.
5. 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)

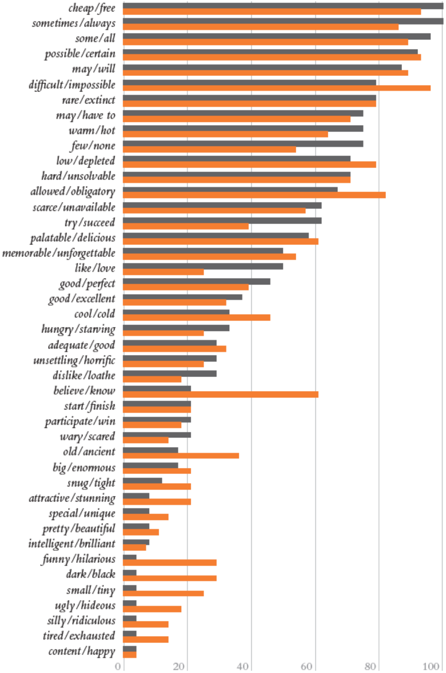
(Grice 1967)

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? —set aside for now

(van Tiel et al. 2016; Sun et al. 2018; Gotzner et al. 2018; Beltrama & Xiang 2013; Ronai & Xiang 2021; Pankratz & van Tiel 2021)

Research goals

Scalar diversity observation: based on **descriptive statistics**, e.g. SI rates range 4%-100%.

Goal 1: provide a **measure to quantify** scalar diversity.

Goal 2: probe what **manipulation can reduce/eliminate** scalar diversity.

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.

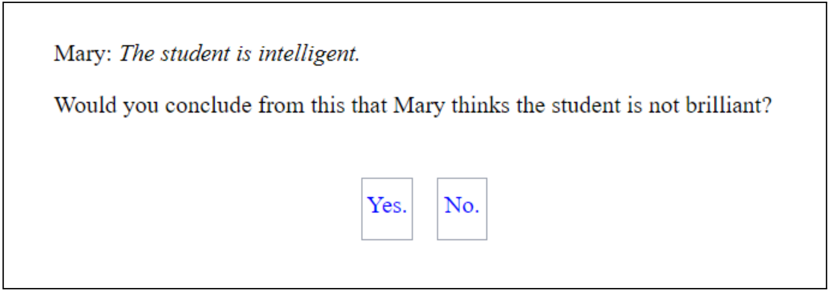
Scale sets from Marneffe & Tonhauser 2019 and van Tiel et al. 2016
+ **COCA searches**: *X or even Y*; *not just X but Y*; *X but not Y* (adjectives, verbs, adverbs).

Filter: semantic tests for asymmetric entailment and cancellability.

Final set: 60 lexical scales.

Experiment 1: replication and semantic manipulation

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



Mary: *The student is intelligent.*

Would you conclude from this that Mary thinks the student is not brilliant?

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

Replication of van Tiel et al. (2016)

Experiment 1: replication and semantic manipulation (*only*)

Inference task: two conditions (between-participants).

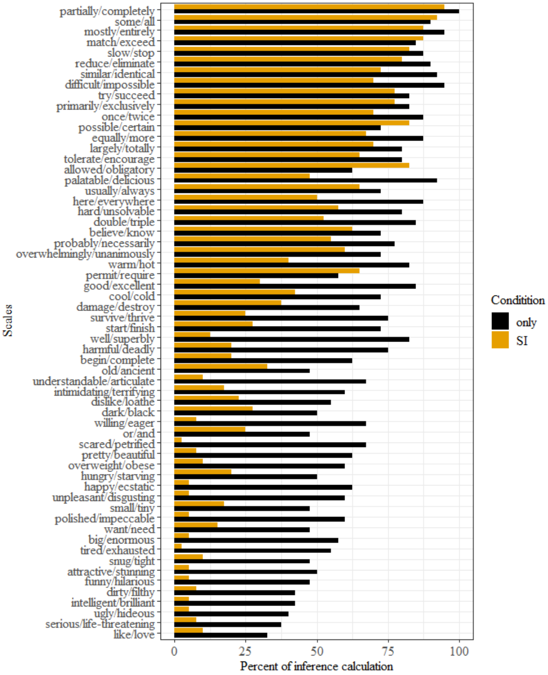
- Bare SI: *The student is intelligent.*
- *Only*: *The student is only intelligent.*

Focus operator *only*: semantically excludes alternatives to the focused element (*intelligent*) (Rooth 1992, 1985).

Predictions:

- *Only*: **higher inference rates**.
  - 100% inference rates, given that it's an entailment, not a cancellable pragmatic inference.
- *Only*: **scalar diversity** will be **reduced (or eliminated)**.

Results



Across the board:

- Higher inference rate with *only* ($p < 0.001$) —though not 100 %.
- Scalar diversity reduced?

Experiment 2: pragmatic manipulation (QUDs)

Questions Under Discussion (**QUDs**, Roberts (1996/2012)):
have an effect on the rate of SI calculation:

(3) A: Did Mary eat all of the cookies?
B: Mary ate some of the cookies.

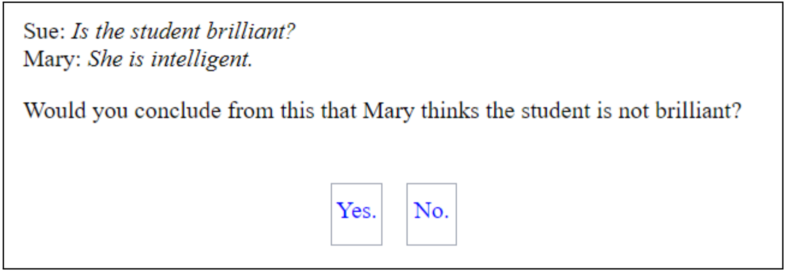
(4) A: Did Mary eat any/some of the cookies?
B: Mary ate some of the cookies.

Higher SI rate in (3) than in (4) (i.a. Cummins & Rohde 2015; Degen & Tanenhaus 2014; Ronai & Xiang 2020; Yang et al. 2018; Zondervan et al. 2008).

We test the effect of such QUD manipulations on scalar diversity.

Experiment 2: pragmatic manipulation (QUDs)

- 40 native speakers of American English; MTurk; IbexFarm.
- Basic inference task identical to Experiment 1.
- Two-condition QUD manipulation: Mary's statement embedded in a dialogue context.
  - Strong-scalar question: *Is the student brilliant?*
  - Weak-scalar question: *Is the student intelligent?*



Sue: *Is the student brilliant?*
Mary: *She is intelligent.*

Would you conclude from this that Mary thinks the student is not brilliant?

Experiment 2: pragmatic manipulation (QUDs)

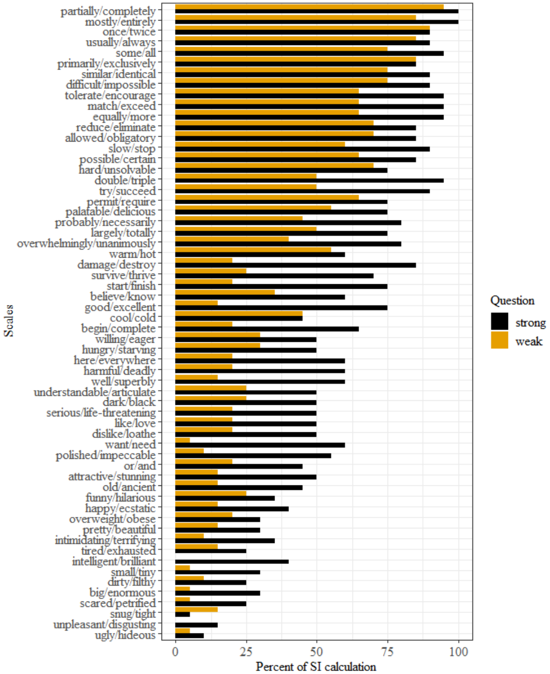
Prediction:

- Strong-scalar question: **higher inference rates** than weak-scalar question.

Hypothesis: scalar diversity is (partly) a consequence of differences in the implicit QUD that scales evoke (Ronai & Xiang 2021).

- Explicit (strong-scalar) question: such differences will be factored out, and **scalar diversity** will be **reduced** (or eliminated).

Results



Across the board:

- Higher inference rate after strong-scalar question ($p < 0.001$).
- Scalar diversity persists?

Taking stock: scalar diversity with semantic/pragmatic manipulation

- Overt exhaustification with *only* and a pragmatic QUD manipulation: both increase inference calculation rates.
- What can way say about whether scalar diversity was reduced?
  Previous work: descriptive statistics.

Quantifying diversity: relative entropy

Treated the normalized **% of “Yes”** responses (i.e. the SI rates) across different scales = **probability distribution**.

Test: does a given SI rate provide enough information to identify the scale that it came from?

Quantifying diversity: relative entropy

Compared each set of SI rates (bare SI, *only*, strong-scalar QUD, weak-scalar QUD) to the **uniform distribution**.

Uniform distribution: each scale leads to the same SI rate.

- The % of “Yes” responses gives 0 information about the identity of the scale it came from.
- Scales cannot be identified by their associated SI rates.

Resulting measure: **relative entropy** (entropy of the uniform distribution minus the entropy of the given SI rates) → quantify how “diverse” the SI rates are.

Quantifying diversity: relative entropy

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:

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

$p(x)$: observed % of "Yes" responses across scales.
$\mathcal{X}$: items.
$q(x) = 1/60$: uniform probability mass function over the 60 scales.

Relative entropy results

Some benchmarks:

- If a set of SI rates is uniform: relative entropy is 0.
- No unique maximal value.
- If we evenly distribute 60 items (=lexical scales) over a 0-100 scale (=SI rates): relative entropy is 0.2912.

Relative entropy measures (as compared to the uniform distribution):

- **Bare SI** (Exp. 1)=0.466 —**substantial difference**, confirming earlier descriptive generalization
- ***Only*** (Exp. 1)=0.046 —**scalar diversity greatly lessened** with focus particle *only*
- **Weak-scalar QUD** (Exp. 2)=0.404 —patterns with **bare SI**
- **Strong-scalar QUD** (Exp. 2)=0.137 —falls in the **middle**

Taking stock

Semantic (overt exhaustification with *only*) and pragmatic (QUD) manipulation are **alike**:

- Lead to **increased inference rate**.

Difference:

- ***Only*** substantially **reduces scalar diversity**.
  - Predicted by how it's not a cancellable pragmatic inference in the first place.
  - Local, semantic cue to reason about alternatives.
- **Strong-scalar QUD** reduces scalar diversity only **to a lesser extent**.
  - Global, pragmatic cue to reason about alternatives.
- Weak-scalar QUD doesn't reduce scalar diversity —neutral baseline.

Conclusions

- We replicate scalar diversity on 60 scales that span grammatical categories.
- First attempt at quantifying how “diverse” the inference rates are, using relative entropy.
- Strong-scalar QUD and focus particle *only* both lead to increased inference rates.
- Semantic manipulation substantially reduces scalar diversity, pragmatic manipulation to a lesser extent.

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.

Cummins & Rohde (2015). Evoking context with contrastive stress: Effects on pragmatic enrichment. Frontiers in Psychology.

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

Degen & Tanenhaus (2014). Processing scalar implicature: A constraint-based approach. Cognitive Science.

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.

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

Roberts (1996/2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. Semantics and Pragmatics.

Ronai & Xiang (2020). Pragmatic inferences are QUD-sensitive: an experimental study. Journal of Linguistics.

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

Rooth (1985). Association with focus. Doctoral Dissertation, UMass Amherst.

Rooth (1992). A theory of focus interpretation. Natural Language Semantics.

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.

Zondervan et al. (2008). Experiments on the role of the question under discussion for ambiguity resolution and implicature computation in adults. Proceedings of SALT 18.

Yang et al. (2018). Context-sensitivity and individual differences in the derivation of scalar implicature. Frontiers in Psychology.