Predicting scalar diversity with context-driven uncertainty over alternatives

Anonymous ACL submission

Abstract

Scalar implicature (SI) arises when a speaker uses an expression (e.g., *some*) that is semantically compatible with a logically stronger alternative on the same scale (e.g., *all*), leading the listener to infer that they did not intend to convey the stronger meaning. Prior work has demonstrated that SI rates are highly variable across scales, raising the question of what factors determine the SI strength for a particular scale. Here, we test the hypothesis that SI rates depend on the listener's confidence in the underlying scale, which we operationalize as uncertainty over the distribution of possible alternatives conditioned on the context. We use a T5 model fine-tuned on a text infilling task to estimate the distribution over context-conditioned alternatives. We measure uncertainty both over the sampled alternatives themselves and over latent clusters among alternatives in sentence embedding space. We find that both scale uncertainty measures predict human SI rates, suggesting that SI depends on listeners' context-driven uncertainty over alternatives.

1 Introduction

Human communication involves not only on the transmission of linguistic signals, but also context-guided inference over the beliefs and goals of other conversational agents (e.g., Sperber and Wilson, 1986; Grice, 1975). One signature pattern of this pragmatic reasoning is scalar implicature (SI). The standard view is that SIs arise as a result of ordered relationships between linguistic items – when a weaker (less informative) item of a scale is uttered, then a listener can infer that the speaker did not have grounds to utter the stronger (more informative) item on that scale. For example, if Alice says "Some of the students passed the exam", Bob can draw the scalar inference that *not all* students passed the exam, even though Alice's utterance would still be semantically true in that scenario.

While this view predicts that SIs are context-independent and generally strong – known as the Homogeneity Assumption (Degen, 2015) – empirical studies have demonstrated a remarkable amount of variance in SI rates both within (Degen, 2015; Li et al., 2021) and across lexical scales (Doran et al., 2009; van Tiel et al., 2016; Gotzner et al., 2018; Pankratz and van Tiel, 2021). This raises the question of what factors determine the SI strength for a particular scale. In a landmark study, van Tiel et al. (2016) test two classes of potential predictors of SI strength: the availability of the strong scalemate given the weak scalemate, and the degree to which scalemates can be distinguished from each other. More recent studies (e.g., Gotzner et al., 2018; Sun et al., 2018; Pankratz and van Tiel, 2021; Ronai and Xiang, 2022) have proposed a variety of other factors such as negative strengthening, polarity, and extremeness, resulting in a collection of manually specified predictors lacking a unified explanation.

Here, we revisit the hypothesis that SI rates depend on the availability of the strong scalemate. While prior work has operationalized availability with measures of the strong scalemate such as word frequency or similarity/association with the weak scalemate (van Tiel et al., 2016; Westera and Boleda, 2020; Ronai and Xiang, 2022), we reframe availability as the listener's *confidence in the underlying scale*. The idea is that upon hearing an utterance with a scalar trigger, listeners must determine the items on the scale as well as the ordering metric before inference can proceed (Hirschberg, 1985). If the listener is less certain about the scale, then they will be less likely to exclude the meaning of any particular strong scalemate.

We operationalize confidence in the scale as uncertainty over the distribution of alternatives that could serve as a strong scalemate to the observed scalar expression. We take the view that an utterance is a suitable alternative vis-à-vis its context-conditioned probability under a predictive language

model optimized on the surface statistics of language. We use T5 (Raffel et al., 2020) fine-tuned on a text infilling task as our practical estimate of the predictive model that humans are using. We evaluate two measures of scale uncertainty: entropy over the alternatives themselves, and entropy over clusters among alternatives in sentence embedding space. We find that both measures predict human SI rates in the dataset collected by van Tiel et al. (2016), suggesting that SI can be explained by listeners' context-driven uncertainty over alternatives.

2 Human data

To obtain human SI strengths, we use the data from Experiment 2 by van Tiel et al. (2016). In our analyses, we only consider the adjectival scales from van Tiel et al.'s original materials, resulting in 32 scales. Each scale is a pair of adjectives ⟨[WEAK], [STRONG]⟩, where the meaning of [STRONG] entails the meaning of [WEAK] (e.g., ⟨*intelligent, brilliant*⟩). The experiment measures whether humans exclude the meaning of [STRONG] upon observing a speaker use [WEAK].

On each trial of the experiment, participants read a prompt of the form "John says: [NP] is [WEAK]", where [WEAK] is an adjective scalar item that may trigger a scalar inference, and [NP] is a noun phrase that sets the context for the scalar item. There were 3 such sentences per scale, which differ from each other only in the NP. For example, the weak scalar item *intelligent* is associated with the sentences "This student/That professor/The assistant is intelligent". Participants were then asked: "Would you conclude from this that, according to John, $[\text{NP}]_{\text{P}}$ is not [STRONG]?", where [STRONG] is the strong scalemate to [WEAK], and $[\text{NP}]_{\text{P}}$ is a pronominalized version of the [NP] in the speaker's original utterance (e.g., "she is not brilliant"). Participants marked their response as Yes or No. The SI rate for a scale is computed as the proportion of Yes responses averaged over participants and sentences.

3 Predictors

We use T5 (Raffel et al., 2020) to estimate all probabilities in our analyses. T5 is a sequence-to-sequence Transformer model (Vaswani et al., 2017) trained to represent language processing tasks as text-to-text problems. We use the pre-trained T5-base model from Huggingface Transformers (Wolf et al., 2020). Since the off-the-shelf T5 model is not optimized for text generation, we use a T5 model that has been fine-tuned on a text infilling task (Qian and Levy, 2022). The model is fine-tuned on a 10-million-token subset of the 2007 portion of the New York Times Corpus (Sandhaus, 2008). The supervision signal is generated by randomly masking some spans of words in a sentence to get the fragmentary context and a plausible completion. At inference time, the model decodes autoregressively via greedy sampling.

3.1 Predictability of strong scalemate

As a baseline, we first consider whether SI rates – i.e., the rate at which [WEAK] is taken to exclude the meaning of [STRONG] – are explained by the context-conditioned predictability of the tested strong scalemate. This is similar to production-based measures of availability, such as the tendency of humans to mention the strong scalemate in a Cloze task (van Tiel et al., 2016; Ronai and Xiang, 2022). However, these metrics are expensive to estimate, especially if we wish to estimate the full distribution of alternatives. We address this by using T5 as a proxy of human predictions.

To measure the predictability of a certain linguistic expression as a strong scalemate under T5, we leverage scalar constructions (Hearst, 1992; de Melo and Bansal, 2013; Pankratz and van Tiel, 2021). Scalar constructions are patterns such as *X, but not Y*, which indicate a scalar relationship between a weak item X and strong item Y. For each weak scalar item in our test materials, we construct a scalar template of the following form:

$$[\text{NP}] \text{ is } [\text{WEAK}], \text{ but not } _____. \quad (1)$$

We have 3 such templates for each scale, where [NP] is given by the 3 sentences from van Tiel et al.'s materials. By embedding the weak scalar item within the *X, but not Y* construction, the model should set up expectations for a potential scalemate in the masked position. For each weak-strong pair from van Tiel et al.'s items, we substitute the strong scalemate into the masked position and compute the surprisal (i.e., negative log probability) at that token under T5.[1] Language model surprisal has been shown to predict psychometric measures of human sentence processing (e.g., Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020), suggesting that the distribution learned by these models captures expectations deployed by humans during real-time language comprehension.

[1] When scalar items are split into multiple tokens, we obtain surprisals by summing over these sub-word tokens.

3.2 Scale uncertainty

Next, we test the hypothesis that SI depends on the listener's uncertainty about the scale implied by the speaker's utterance. Depending on the context, a single word (e.g., *bad*) could lie on multiple scales – e.g., "The food is bad" might imply that the food is not rotten, whereas "The score is bad" might imply that the score is not terrible. This uncertainty is not a function of a particular scalemate (unlike the availability measure described in Section 3.1 and in prior work), but rather a property of the scalar trigger and the context in which it is observed.

We operationalize scale uncertainty as uncertainty over the distribution of possible alternatives conditioned on the context. To obtain a set of candidate alternatives A, we sample $N = 100$ completions from the T5 infilling model given the scalar template in Equation (1).[2] During decoding, we restrict the maximum number of generated tokens to 5, and only keep the unique completions. After obtaining A, we operationalize scale uncertainty in two ways: string-based and hierarchical.

String-based. Our first measure of scale uncertainty is uncertainty over the strings that fill the masked position in the scalar template (Equation (1)). That is, we normalize the surprisals of each $a \in A$ to obtain a probability distribution over alternatives, and then compute the Shannon entropy over this distribution. Our hypothesis predicts that lower entropy reflects lower uncertainty over the underlying scale, and thus results in a stronger SI.

This method implicitly assumes that surface-level linguistic forms (i.e., strings) are the alternatives driving scalar inferences. As a single concept can be expressed with multiple forms, however, the surprisal over forms may not be a good estimate of the surprisal of the underlying concept.

Hierarchical. An alternate view is that listeners do not reason about alternatives at the level of linguistic forms (i.e., strings), but instead a deeper conceptual level (Buccola et al., 2021). To formalize this idea, we treat scales as latent classes that may give rise to multiple alternative strings, and measure scale uncertainty as uncertainty over these classes. This motivates using hierarchical methods to identify clusters among alternatives in some conceptual representation. Using each $a \in A$, we substitute a into the masked position in the scalar template (Equation (1)) to obtain a full sentence. We then use Sentence-T5 (Ni et al., 2021) to obtain a 768-dimensional embedding of the entire sentence as a proxy of a conceptual representation. We assume sentences close in this space are more likely to reflect the same underlying scale, and distant sentences are likely to reflect different scales.

To identify clusters among sentence embeddings, we use hierarchical agglomerative clustering (HAC), which recursively merges pairs of clusters using Euclidean distance and minimizes the variance of the clusters being merged at each step. We do not specify a fixed number of clusters, but instead set a distance threshold above which clusters will not be merged.[3] After running the HAC algorithm on the alternative embeddings for each weak scalemate, we obtain a score for each cluster by summing over the probabilities assigned by T5 to each alternative within that cluster. We then normalize these by-cluster scores and compute the Shannon entropy of this distribution. As with the string-based approach, our hypothesis predicts that lower entropy should result in a stronger SI.

4 Results

We computed the metrics described in Section 3 on the data from Experiment 2 of van Tiel et al. (2016), and evaluated the ability of each metric to predict the diversity of scalar inference rates.

We first consider the baseline measure of availability; i.e., the in-context predictability of the strong scalemate (Figure 1a). Each point represents a scale, with values averaged over the 3 sentence templates presented in van Tiel et al.'s Experiment 2. We find no relationship between strong scalemate surprisal and SI rate (Pearson $\rho = -0.085$, $p = 0.643$), in line with van Tiel et al.'s original finding that availability is not predictive of SI rate.

Next, we turn to scale uncertainty. Figure 1b shows the results for the string-based measure of scale uncertainty; i.e., the entropy over completions sampled from T5 in a scalar construction. We find a significant negative correlation between entropy over strings and human SI rate, as predicted (Pearson $\rho = -0.501$, $p = 0.003$). However, we note that a large portion of the data is in the high-entropy region, likely because most alternatives are low-probability. Figure 1c shows the results for the

[2] The completions are not guaranteed to be scalar items, but we take this to be a first approximation. All results are averaged over 4 random seeds for the sampling of alternatives.

[3] We set the distance threshold to 0.8 based on qualitative evaluations of pilot experiments.

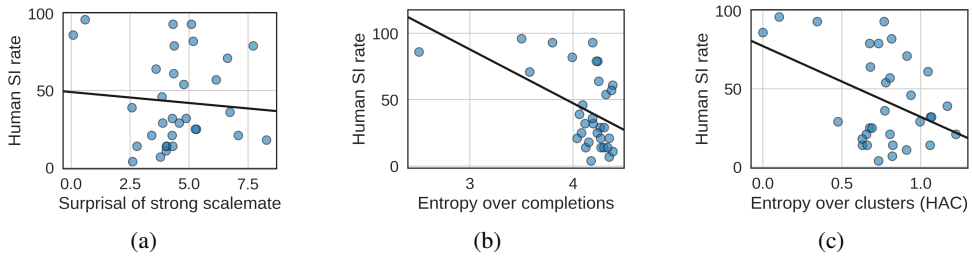

Figure 1: Best-fit linear relationship between human SI rates (van Tiel et al., 2016) and three predictors (Section 3): (a) Surprisal of strong scalemate under T5. (b) Entropy over alternative strings sampled by T5 in scalar construction. (c) Entropy over probabilities of clusters identified by hierarchical agglomerative clustering.

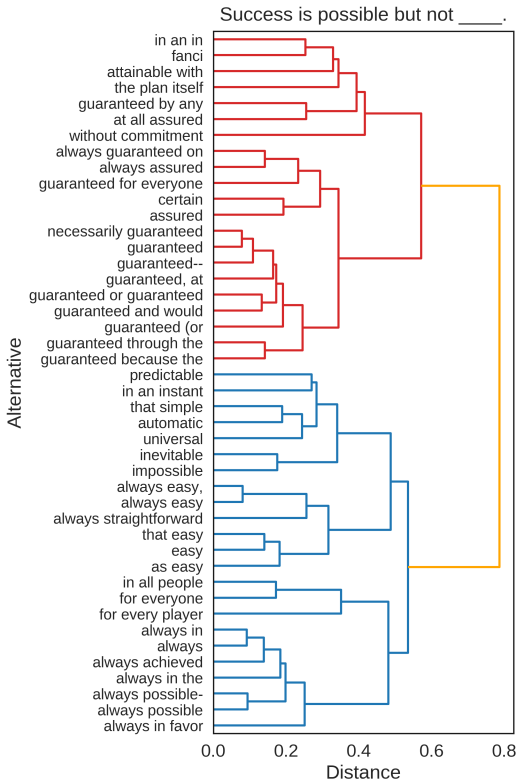

Figure 2: Example of clusters identified by HAC model among alternatives in sentence embedding space.

HAC model, which measures uncertainty over clusters among the alternatives in sentence embedding space. Again, we find a significant negative correlation between entropy over clusters and human SI rate (Pearson $\rho = -0.417, p = 0.012$).

Finally, we examined the clusters identified by the HAC model. Figure 2 shows an example dendrogram for the template "Success is possible but not ____.", for which the model identifies 2 clusters of alternatives to the weak scalar item *possible*. One cluster (red) contains *certain*, the strong scalemate tested in van Tiel et al.'s experiments, as well as semantically similar alternatives such as "assured" and "guaranteed". The other cluster (blue) contains alternatives such as "in an instant", "automatic", "(always) easy", and "always straightforward", capturing the potential of the weak scalar item *possible* to lie on a scale like ⟨*possible, easy*⟩.

5 Discussion

We tested the hypothesis that SI rates depend on the listener's confidence in the underlying scale, using T5 as our estimate of the predictive distribution used by humans. Using data from a previously conducted experiment (van Tiel et al., 2016), we found that scale uncertainty was a significant predictor of SI rates: on average, when uncertainty (i.e., entropy over sampled alternatives, or over clusters of alternatives in sentence embedding space) is lower, humans are more likely to draw a scalar inference.

An open question is why scale uncertainty predicts SI rates, while strong scalemate surprisal and the original availability measures from van Tiel et al. (2016) are poor predictors. Returning to the example illustrated by Figure 2, if *guaranteed* has low surprisal when the weak item is *possible*, then entropy might be low but the surprisal of *certain* could be high. Thus, when many of the generated alternatives are similar in meaning, a high-surprisal scalemate does not necessarily reflect high surprisal about the underlying scale itself. Another possibility is that T5's distribution is not well-tuned to the simple sentential contexts used in van Tiel et al.'s (and other researchers') studies of scalar diversity.

Looking forward, our methods can be applied to scales that are ordered by ad-hoc relationships instead of entailment (Hirschberg, 1985). Beyond predicting scalar diversity, our approach suggests a way to derive quantitative behavioral predictions from non-linguistic alternatives (Buccola et al., 2021), and supports the idea that context-driven expectations may give rise to pragmatic behaviors.

References

Brian Buccola, Manuel Križ, and Emmanuel Chemla. 2021. Conceptual alternatives: Competition in language and beyond. *Linguistics and Philosophy*.

Gerard de Melo and Mohit Bansal. 2013. Good, Great, Excellent: Global Inference of Semantic Intensities. *Transactions of the Association for Computational Linguistics*, 1:279–290.

Judith Degen. 2015. Investigating the distribution of *some* (but not *all*) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 8(11):1–55.

Ryan Doran, Rachel E. Baker, Yaron McNabb, Meredith Larson, and Gregory Ward. 2009. On the Non-Unified Nature of Scalar Implicature: An Empirical Investigation. *International Review of Pragmatics*, 1(2):211 – 248. Place: Leiden, The Netherlands Publisher: Brill.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

Nicole Gotzner, Stephanie Solt, and Anton Benz. 2018. Scalar Diversity, Negative Strengthening, and Adjectival Semantics. *Frontiers in Psychology*, 9:1659.

Herbert P. Grice. 1975. Logic and Conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Speech Acts*, volume 3, pages 41–58. Academic Press.

Marti A. Hearst. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Julia Bell Hirschberg. 1985. *A Theory of Scalar Implicature (Natural Languages, Pragmatics, Inference)*. PhD Thesis, University of Pennsylvania.

Elissa Li, Sebastian Schuster, and Judith Degen. 2021. Predicting Scalar Inferences From "Or" to "Not Both" Using Neural Sentence Encoders. In *Proceedings of the Society for Computation in Linguistics*, volume 4.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models.

Elizabeth Pankratz and Bob van Tiel. 2021. The role of relevance for scalar diversity: a usage-based approach. *Language and Cognition*, 13(4):562–594. Edition: 2021/08/16 Publisher: Cambridge University Press.

Peng Qian and Roger Levy. 2022. Flexible generation from fragmentary linguistic input. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. To appear.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Eszter Ronai and Ming Xiang. 2022. Three factors in explaining scalar diversity. In *Proceedings of Sinn und Bedeutung 26*.

Evan Sandhaus. 2008. The New York Times Annotated Corpus (LDC2008T19).

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302 – 319.

Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and Cognition*. Wiley-Blackwell.

Chao Sun, Ye Tian, and Richard Breheny. 2018. A Link Between Local Enrichment and Scalar Diversity. *Frontiers in Psychology*, 9:2092.

Bob van Tiel, Emiel van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar Diversity. *Journal of Semantics*, 33(1):137–175.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Matthijs Westera and Gemma Boleda. 2020. A closer look at scalar diversity using contextualized semantic similarity. *Proceedings of Sinn und Bedeutung*, 24(2):439–454.

Ethan Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the Cognitive Science Society*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.