Expectations over Unspoken Alternatives Predict Pragmatic Inferences

Jennifer Hu[◇], **Roger Levy**[◇], **Judith Degen**[†], **Sebastian Schuster**[‡]

[◇]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, USA
{jennhu, rplevy}@mit.edu
[†]Department of Linguistics, Stanford University, USA
jdegen@stanford.edu
[‡]Department of Language Science and Technology, Saarland University, Germany
seschust@lst.uni-saarland.de

Abstract

Scalar inferences (SI) are a signature example of how humans interpret language based on unspoken alternatives. While empirical studies have demonstrated that human SI rates are highly variable—both within instances of a single scale, and across different scales—there have been few proposals that quantitatively explain both cross- and within-scale variation. Furthermore, while it is generally assumed that SIs arise through reasoning about unspoken alternatives, it remains debated whether humans reason about alternatives as linguistic forms, or at the level of concepts. Here, we test a shared mechanism explaining SI rates within and across scales: context-driven expectations about the unspoken alternatives. Using neural language models to approximate human predictive distributions, we find that SI rates are captured by the expectedness of the strong scalemate as an alternative. Crucially, however, expectedness robustly predicts cross-scale variation only under a meaning-based view of alternatives. Our results suggest that pragmatic inferences arise from context-driven expectations over alternatives, and these expectations operate at the level of concepts.

1 Introduction

Much of the richness of linguistic meaning arises from what is left unsaid (e.g., Grice, 1975; Sperber and Wilson, 1986; Horn, 1989). For example, if Alice says ‘‘Some of the students passed the exam’’, Bob can infer that Alice means *not all* students passed the exam, even though Alice’s utterance would still be logically true if all students had passed. One explanation of this inference is that Bob reasons about the unspoken **alternatives** that were available to the speaker. Under the assumptions that (1) speakers generally try to be informative, (2) Alice has full knowledge of the situation, and (3) it would have been relevant and more informative for Alice to say ‘‘All of the students passed the exam’’, Alice’s choice to say ‘‘some’’ suggests that she believes the sentence with ‘‘all’’ is false. This inference pattern is more generally known as **scalar inference** (SI), which arises from orderings between linguistic items (scales).

SI has often been treated as a categorical phenomenon: When a speaker utters a weaker (less informative) item on a scale, a listener rules out the meaning of stronger (more informative) items on that scale (e.g., Levinson, 2000). However, empirical studies have demonstrated substantial variability in the rates at which humans draw SIs, both within instances of a single scale (Degen, 2015; Eiteljoerge et al., 2018; Li et al., 2021) and across scales formed by different lexical items (e.g., Doran et al., 2009; Beltrama and Xiang, 2013; van Tiel et al., 2016; Gotzner et al., 2018; Pankratz and van Tiel, 2021; Ronai and Xiang, 2022). For example, consider the following instances of the scale ⟨*some*, *all*⟩:

(1) a. I like some country music.
b. I like some, but not all, country music.

(2) a. It would certainly help them to appreciate some of the things that we have here.
b. It would certainly help them to appreciate some, but not all, of the things that we have here.

Degen (2015) finds that humans are highly likely to consider (1-a) as conveying a similar meaning as (1-b), but unlikely to consider (2-a) as conveying a similar meaning as (2-b) (Figure 1a). Similarly, consider the following instances of the scales

Code and data can be found at: https://github.com/jennhu/expectations-over-alternatives.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 885–901, 2023. https://doi.org/10.1162/tacl_a_00579
Action Editor: Ehud Reiter. Submission batch: 11/2022; Revision batch: 3/2023; Published 7/2023.

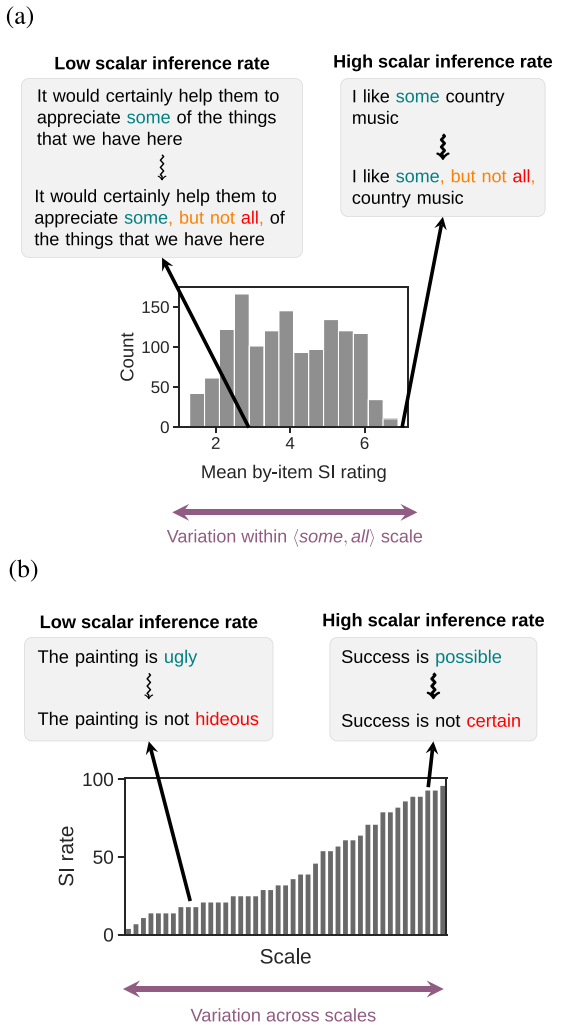

Figure 1: (a) Distribution of human scalar inference (SI) ratings (on scale of 1–7) across instances of the ⟨*some*, *all*⟩ scale (reproduction of Figure 1, Degen, 2015). (b) Average SI rates across scales formed by different lexical items (reproduction of Figure 2, van Tiel et al., 2016).

⟨*possible*, *certain*⟩ and ⟨*ugly*, *hideous*⟩, which both consist of adjectives ordered by entailment:

(3) a. Success is possible.
b. Success is not certain.

(4) a. The painting is ugly.
b. The painting is not hideous.

van Tiel et al. (2016) find that humans are highly likely to conclude that (3-a) implies (3-b), but unlikely to conclude that (4-a) implies (4-b) (Figure 1b).

While cross-scale and within-scale variation have typically been studied as distinct empirical phenomena, they both reflect gradedness in listener inferences based on alternatives and context. It therefore seems desirable to explain these empirical findings with a shared account, but there have been few proposals that quantitatively explain both within- and cross-scale variation. For example, cross-scale variation can be explained by intrinsic properties of the scale (e.g., whether the strong scalemate refers to an extreme endpoint; van Tiel et al., 2016), but these factors cannot explain variation within instances of a single scale. On the other hand, many factors explaining within-scale variance are scale-specific (e.g., the partitive ‘‘of the’’ for ⟨*some*, *all*⟩; Degen, 2015) and may not generalize to new scales.

Here, we investigate a shared account of SI rates within and across scales. Since the alternatives are not explicitly produced (by definition), the listener has uncertainty over which alternatives the speaker could have used—and therefore, which strong scalemates ought to be negated through SI. Building upon constraint-based accounts of human language processing (Degen and Tanenhaus, 2015, 2016), we test the hypothesis that SIs depend on the availability of alternatives, which depend on context-driven expectations maintained by the listener. For example, if a speaker says ‘‘The movie was good’’, the listener might predict that *amazing* is a more likely alternative than *funny* to the weak term *good*. An expectation-based view predicts that the listener would be thus be more likely to infer that the movie is not amazing (according to the speaker), and less likely to infer that the movie is not funny. However, while Degen and Tanenhaus (2015, 2016) have argued that listeners maintain context-driven expectations over alternatives, these studies have primarily investigated a single scale (⟨*some*, *all*⟩) in small domains, arguing from qualitative patterns and in the absence of a formal theory.

Furthermore, while it is generally assumed that SIs arise based on reasoning about unspoken alternatives, it remains debated whether humans reason about alternatives as linguistic structures (e.g., Katzir, 2007; Fox and Katzir, 2011), or at the level of concepts (e.g., Gazdar, 1979; Buccola et al., 2021). Returning to the earlier example, if the weak scalemate is *good*, listeners may reason about a concept like VeryGood instead of a specific linguistic expression like *amazing*. In this sense, the listener’s uncertainty about alternatives might arise from uncertainty about

Dataset	Type of variation	# participants	# scales	# contexts per scale	# data points per item
Degen (2015)	Within-scale	243	1	1363	~ 10
Ronai and Xiang (2022)	Cross-scale	40	57	1	40
Pankratz and van Tiel (2021)	Cross-scale	1970	50	1	~ 40
Gotzner et al. (2018)	Cross-scale	220	67	1	40
van Tiel et al. (2016)	Cross-scale	28	39	3	10

Table 1: Details of human data used in our analyses. An item is a unique (scale, context) combination.

both the scale itself (*Is the speaker implying the plot wasn't amazing, or that the jokes weren't funny?*), as well as the exact word forms under consideration by the speaker (*Is the speaker implying the movie wasn't amazing, fantastic, or wonderful?*). Despite theoretical debates about the nature of alternatives, however, the role of concept-based alternatives in SI has not been tested in a systematic, quantitative way.

We provide a formalization of an expectation-based account of alternatives and test it on both string-based and concept-based views of alternatives. Instead of empirically estimating human expectations over alternatives (cf. Ronai and Xiang, 2022), we use neural language models as an approximation, which allows us to generate predictions for arbitrary sentences and contexts. We test the account's predictions on human SI rates within the ⟨*some*, *all*⟩ scale (Degen, 2015), and across 148 scales from four datasets (van Tiel et al., 2016; Gotzner et al., 2018; Pankratz and van Tiel, 2021; Ronai and Xiang, 2022). We find support for the expectation-based account, and also provide the first evidence that concept-based alternatives may be underlying a wide range of SIs. Our results suggest that pragmatic inferences may arise from context-driven expectations over unspoken alternatives, and these expectations operate at the level of concepts.

2 Background

2.1 Within-scale Variation

Within-scale variation refers to the variation in SI rates across instances of a single scale, such as ⟨*some*, *all*⟩. To explore SI variation within the scale ⟨*some*, *all*⟩, we use the dataset collected by Degen (2015), which features 1363 naturalistic sentences containing a ''some''-NP from the Switchboard corpus of telephone dialogues (Godfrey et al., 1992) (Table 1). For each sentence, SI rates were measured using a sentence-similarity paradigm. On each trial, participants saw two sentence variants: the original sentence containing ''some'', and a minimally differing sentence where '', but not all,'' was inserted directly after ''some''. Participants were asked, ''How similar is the statement with 'some, but not all' to the statement with 'some'?'' and indicated responses (similarity judgments) on a seven-point Likert scale. If the speaker's originally intended meaning clearly includes an implicature, then making the implicature explicit by inserting '', but not all,'' should not change the meaning of the sentence, so similarity judgments should be high. Thus, a higher similarity judgment indicates a stronger SI.

Degen (2015) finds substantial variation in SI rates across contexts, challenging the idea that the ''some, but not all'' inference arises reliably without sensitivity to context (Horn, 1989; Levinson, 2000). She also reports several features that predict SI rates, such as whether ''some'' occurs with the partitive ''of the'', or whether the ''some''-NP is in subject position. However, these features may be highly specific to the ⟨*some*, *all*⟩ scale, and it is unclear whether a more general mechanism may also explain variation within or across other scales.

2.2 Cross-scale Variation (Scalar Diversity)

Cross-scale variation refers to the variation in SI rates across scales formed by different lexical items. To explore this, we use SI rates across 148 unique scales from four datasets, summarized in Table 1. Each scale involves a pair of English words (adjectives, adverbs, or verbs) of the form ⟨`[WEAK]`, `[STRONG]`⟩, where `[WEAK]` is less informative than `[STRONG]` (e.g., ⟨*intelligent*, *brilliant*⟩).[1] For each dataset, SI rates were measured through a binary choice task. Participants saw a character make a short, unembedded statement consisting of a simple noun phrase subject and a predicate with a weak scalar item (e.g., ''John says: This student is intelligent.''). Their

[1]We excluded scales where one of the items was formed by a multi-word expression (e.g., ⟨*may*, *have to*⟩).

task was to indicate (*Yes* or *No*) whether they would conclude that the speaker believes the negation of a strong scalar item (e.g., ‘‘Would you conclude from this that, according to John, she is not brilliant?’’). The SI rate for a scale is the proportion of *Yes* responses.

This method has revealed large variation in SI rates, ranging from 4% (⟨*ugly*, *hideous*⟩) to 100% (⟨*sometimes*, *always*⟩) (van Tiel et al., 2016). van Tiel et al. (2016) test two classes of factors that might predict SI rates: the availability of the strong scalemate given the weak scalemate, and the degree to which scalemates can be distinguished from each other. They find SI rates are predicted by measures of scalemate distinctness (e.g., whether the strong scalemate forms a fixed endpoint on the scale), but not by availability (but see Westera and Boleda, 2020; Ronai and Xiang, 2022). Other studies have proposed additional scale-intrinsic factors (e.g., Gotzner et al., 2018; Sun et al., 2018; Pankratz and van Tiel, 2021). However, structural properties of a scale cannot explain variablity in SI rates *within* a scale, as these properties do not change across contexts.

While others have proposed context-dependent factors—which could, in principle, explain both cross- and within-scale variation—these factors often lack explanatory power in practice. For example, Ronai and Xiang (2021) find that the prominence of the Question Under Discussion (Roberts, 2012) is correlated with SI rates, but only for unbounded scales (i.e., scales where neither scalemate has a fixed, extreme meaning).

3 An Expectation-based Account of SI

Theoretically, it is the set of alternative utterances—utterances that the speaker could have used, but didn’t—that drive scalar implicature, and in principle every possible utterance in a language might be an alternative to every other. However, at an algorithmic level (Marr, 1982), it would be intractable for listeners to perform inference over this entire set. Furthermore, the signature pattern of SI would not arise without restrictions on the alternatives: otherwise, ‘‘[WEAK], but not [STRONG]’’ and ‘‘[STRONG]’’ would both be alternatives to ‘‘[WEAK]’’, leading to contradictory inferences without a mechanism for breaking symmetry (Kroch, 1972; Katzir, 2007; Breheny et al., 2018).

To solve this symmetry problem, some approaches restrict alternatives based on structural complexity through grammar-internal mechanisms (e.g., Katzir, 2007; Fox and Katzir, 2011). However, these theories do not capture the uncertainty that listeners maintain, and are difficult to test quantitatively. Here, we test the view that listeners form probabilistic expectations over alternatives, given information from their interaction with the speaker. In the remainder of this section, we first discuss the conceptual predictions of an expectation-based account of SI, and then describe how we operationalize these predictions using neural language models.

Suppose that a listener hears a sentence with a weak scalar term [WEAK] (e.g., ‘‘This student is intelligent’’). To rule out the meaning of a particular strong scalemate [STRONG] (e.g., the student is not *brilliant*), the listener must have reason to believe that the speaker would have said [STRONG] if they had intended to convey the strong meaning. However, since the alternatives are not explicitly produced, the listener has some degree of uncertainty over what alternatives were considered by the speaker. If it is likely that the speaker would have said [STRONG] to convey the strong meaning, then their choice to say [WEAK] suggests that they did not have grounds to say [STRONG]—and thus, an SI should be more likely to arise.

The key question, then, is how listeners estimate which alternatives are likely to be considered by the speaker. An expectation-based account proposes that listeners integrate contextual and grammatical cues to maintain probabilistic expectations over these alternatives. A scalemate that is more probable (given these cues) should be more likely to enter the scalar inference computation. Thus, this account predicts that the more expected the strong scalemate is as an alternative to the weak scalemate, the higher SI rates should be.

3.1 String-based View of Alternatives

When an alternative is likely to be a strong scalemate, listeners should be more likely to rule out its meaning, resulting in higher SI rates. Conditioned on the context and the speaker’s choice to use [WEAK], the listener must estimate the probability of [WEAK] and [STRONG] being contrasted in a scalar relationship. Since it is difficult to directly estimate this probability, we construct a sentence

frame where the probability of [STRONG]—at the level of forms—approximates the probability of [STRONG] being in a scalar relationship with a weak scalemate [WEAK]. This approach allows us to re-frame the problem of estimating listeners' expectations over strong scalemates into a word prediction problem.

To do this, we use the scalar construction ''*X, but not Y*'', which in many cases suggests that *Y* is a strong scalemate to *X* (Hearst, 1992; de Melo and Bansal, 2013; van Miltenburg, 2015; Pankratz and van Tiel, 2021). For a given utterance [CONTEXT] [WEAK] [CONTEXT] and hypothesized scale ⟨[WEAK], [STRONG]⟩, we form a sentence that explicitly states the SI:

[CONTEXT] [WEAK], but not [STRONG], [CONTEXT]

scalar construction

(1)

To test how expected [STRONG] is as an alternative to [WEAK], we need to estimate how likely a human would predict [STRONG] to appear in the [STRONG] position in (1).[2] Instead of attempting to directly measure these predictions (cf. Ronai and Xiang, 2022, see (3)), we approximate this with neural language models. We measure how unexpected [STRONG] is by computing its surprisal (negative log probability) under a language model, conditioned on the rest of the sentence. Since surprisal measures *un*expectedness, we predict a negative relationship between SI rate and the surprisal of the strong scalemate.

This predictor is closely related to the notion of an SI's ''relevance'' (Pankratz and van Tiel, 2021). Under usage-based theories of language (e.g., Tomasello, 2003; Bybee and Beckner, 2015), if a weak scalar term is encountered frequently in a scalar relationship with a particular strong term, then the scalar relationship between these items will be enforced. Thus, Pankratz and van Tiel (2021) measure the relevance of an SI by counting corpus frequencies of the scalemates in the string ''[WEAK], but not [STRONG]''. This is conceptually aligned with our setup, where we might expect higher corpus frequencies to correspond to lower surprisal under a language model. However, our predictor differs from Pankratz and van Tiel's in an important way: they aim to measure the ''general relevance'' of an SI, which they define as ''relevance even in the absence of a situated context.'' It is unclear how general relevance can explain variation in SI rates within instances of a scale. By using context-conditioned probabilities from a language model, our predictor could account for both the general frequency of ''[WEAK], but not [STRONG]'' as well as expectations driven by the context in which the scale occurs.

3.2 Concept-based View of Alternatives

The method described above implicitly treats linguistic forms as the alternatives driving scalar inferences. However, recent proposals have advanced the view that alternatives are not linguistic objects, but instead operate at the level of more general reasoning preferences (Buccola et al., 2021). On this view, alternatives are constructed by replacing primitives of the concept expressed by the speaker with primitives of equal or less complexity.

Here, we test a generalization of this concept-based view of alternatives. Suppose, for example, a speaker uses the weak scalar term *big*. On a concept-based view, the listener may infer that the speaker is contrasting *big* with a concept like VeryBig instead of a particular linguistic expression like *enormous*. However, in the experiments mentioned in Section 2.2, the SI process likely needs to be grounded in linguistic forms before the listener makes a judgment about a particular strong scalemate (in string form). One hypothesis is that upon hearing an expression with a weak scalemate, a stronger conceptual alternative is activated, which in turn probabilistically activates all the strings that could reflect it. Returning to our earlier example, if the conceptual alternative is VeryBig, and *huge*, *massive*, and *enormous* are string-based realizations of that alternative, they may be assigned a high likelihood. When asked about a specific string-form alternative (e.g., ''The elephant is big. Would you conclude

[2] Another approach would be to measure the expectedness of [STRONG] in the template [CONTEXT] [STRONG] [CONTEXT]—that is, by replacing [WEAK] with [STRONG] in the speaker's original utterance. This template would instantiate the theory that listeners determine alternatives based on the context. In contrast, the template we use in (1) instantiates the theory that listeners form expectations over alternatives based on the context as well as the speaker's usage of [WEAK]. We return to this topic in Section 7.1.

that it is not enormous?''), humans may endorse the SI if the probability of conceptually similar linguistic alternatives is sufficiently high, even if the probability of the tested alternative (here, *enormous*) is low.

If SIs involve reasoning about conceptual alternatives, then surprisal values estimated from assumed string-form alternatives may be poor estimates of the true relevant surprisal, as a single concept could be expressed with multiple forms. Therefore, in addition to assessing whether expectedness of specific linguistic forms predicts SI rates (Section 3.1), we also test a second predictor which approximates the expectedness of conceptual alternatives. To do this, we need a set of alternatives $\mathcal{A}$ that could serve as potential linguistic scalemates. As described in more detail in Section 4.3 and 5.3, we obtain $\mathcal{A}$ by taking a fixed set of words with the same part of speech as the weak scalemate, inspired by grammatical theories of alternatives (e.g., Rooth, 1985; Katzir, 2007).[3]

Using this alternative set $\mathcal{A}$, we compute the weighted average surprisal of $\mathcal{A}$ using weights determined by the conceptual similarity between each alternative and the tested strong scalemate. We use GloVe embeddings (Pennington et al., 2014) as an approximation for conceptual representations of scalar items, and cosine similarity between GloVe vectors to approximate conceptual similarity.

For each scale $\langle$ [WEAK], [STRONG] $\rangle$, we obtain weights by computing the cosine similarity between the GloVe embeddings for [STRONG] ($v_{\texttt{[STRONG]}}$) and each potential alternative a (v_a) in the alternative set $\mathcal{A}$. We compute the weighted average probability over $\mathcal{A}$ using these weights, and then take the negative log to obtain the weighted average surprisal:

$$-\log\left(\frac{\sum_{a\in\mathcal{A}} P(a)\cdot \text{cossim}(v_{\texttt{[STRONG]}}, v_a)}{\sum_{a\in\mathcal{A}} \text{cossim}(v_{\texttt{[STRONG]}}, v_a)}\right) \quad (2)$$

If there are many conceptually similar alternatives with low surprisal, then the weighted average surprisal will be low, even if the surprisal of the tested scalemate is high. Therefore, weighted average suprisal forms a proxy for concept-based surprisal, which we compare to string-based suprisal.

[3]We adopt a liberal view of alternatives to avoid undergeneration. However, an important open question is how alternatives are determined, which we leave for future work.

4 Predicting Variation Within ⟨*some*, *all*⟩

4.1 Human Data

To investigate variation within the scale ⟨*some*, *all*⟩, we use human SI strength ratings collected by Degen (2015). These ratings were measured by asking participants to rate the similarity (1–7) between a sentence with ''some'' and a minimally differing sentence with ''some, but not all''. See Section 2.1 for details.

4.2 Model

Following the experiment conducted by Degen (2015), we construct scalar templates by inserting '', but not all,'' after the occurrence of ''some'' in each sentence from the dataset. Since this scalar construction (''some, but not all,'') often occurs in the middle of the sentence, we use the bidirectional language model BERT (Devlin et al., 2019) to measure model expectations at the position of the strong scalemate. Concretely, we replace ''all'' with the [MASK] token and measure BERT's probability distribution at that token. All models in our study are accessed via the Huggingface `transformers` library (Wolf et al., 2020).

4.3 Candidate Alternatives

For our string-based surprisal predictor (Section 3.1), we are only concerned with the surprisal of the alternative *all* in the [STRONG] position in (1). However, to compute our concept-based surprisal predictor Section 3.2), we need a set of candidate alternatives that could potentially serve as the strong scalemates implied by the speaker. Since the alternatives to *some* are highly constrained by the grammar, we manually constructed a set of English quantifiers that can be used in contrast to *some*: *each*, *every*, *few*, *half*, *much*, *many*, *most*, and *all*.

4.4 Results

Figure 2 shows the relationship between our predictors and human SI ratings for Degen's (2015) dataset of variation within ⟨*some*, *all*⟩. We find that both string-based and concept-based surprisal are indeed negatively correlated with human similarity judgments (string-based: Figure 2a,

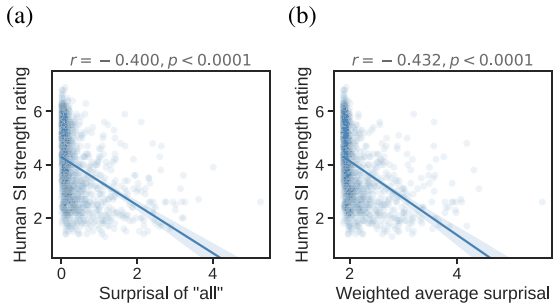

Figure 2: Relationship between human SI strength ratings within ⟨*some*, *all*⟩ scale (Degen, 2015) and BERT-derived predictors: (a) surprisal of scalemate *all* in the scalar construction, and (b) weighted average surprisal over the full set of candidate alternatives (Section 4.3). Each point represents a sentence. Shaded region denotes 95% CI.

Predictor	β	p
DEGEN (2015) PREDICTORS		
Partitive	0.658	< 0.0001
Strength	-0.470	< 0.0001
Mention	0.287	< 0.0001
Subjecthood	0.495	< 0.0001
Modification	0.157	< 0.01
Log sentence length	0.189	< 0.0001
OUR PREDICTORS		
String-based surprisal	**0.008**	**0.960**
Concept-based surprisal	$\mathbf{-0.782}$	$\mathbf{< 0.001}$

Table 2: Summary of the full regression model, including original predictors from Degen (2015) (see the original study for a detailed description of each of the predictors).

Pearson $\rho = -0.400, p < 0.0001$; concept-based: Figure 2b, $\rho = -0.432, p < 0.0001$).[4]

We additionally conducted a multivariate analysis including our two new predictors (string- and concept-based surprisal) among the predictors investigated in Degen's original study. We centered and transformed all variables according to Degen's original analyses. The results are summarized in Table 2. We find that the original predictors remain statistically significant, and that concept-based surprisal (but not string-based surprisal) is a significant predictor in the full model. This suggests that listeners draw stronger scalar inferences when *all*—or a conceptually similar alternative—is more expected in a given context.

5 Predicting Variation Across Scales

5.1 Human Data

To investigate variation across scales, we use human SI rates collected by four studies (Ronai and Xiang, 2022; Pankratz and van Tiel, 2021; Gotzner et al., 2018; van Tiel et al., 2016). SI rates were measured by showing participants a sentence with the weak scalemate (e.g., ''The student is intelligent''), and asking whether they would endorse the negation of the strong scalemate (e.g., ''The student is not brilliant''). See Section 2.2 for details.

5.2 Model

We construct scalar templates following the pattern summarized in Table 3. Since in each case the strong scalemate is the final word in the sentence,[5] we use an autoregressive language model to measure expectations over potential scalemates in the `[STRONG]` position. We use the base GPT-2 model (Radford et al., 2019) via HuggingFace and obtain model surprisals through the SyntaxGym command-line interface (Gauthier et al., 2020).

5.3 Candidate Alternatives

Recall from Section 3.2 that we need a set of potential linguistic alternatives to compute the weighted average surprisal. We take this set of alternatives to be a set of words with the same part of speech (POS) as the weak scalemate and obtain these candidate alternative sets by extracting lists of English adjectives, adverbs, and verbs from WordNet (Miller, 1995). We then used NLTK (Loper and Bird, 2002) to find the words satisfying finer-grained POS tags (`JJ` for adjectives, `RB` for adverbs, and `VB` for verbs), and sorted each POS set according to word frequencies from the OpenSubtitles corpus (Lison and Tiedemann, 2016).[6,7]

[4] We note that the relationship between surprisal and SI ratings visually appears highly non-linear in Figure 2. We expect this is because the scalemate *all* is highly expected in most contexts, so the surprisal values of *all* are concentrated at low values. There is a stronger linear relationship between SI ratings and raw probabilities (string-based: $\rho = 0.482, p < 0.0001$; concept-based: $\rho = 0.513, p < 0.0001$).

[5] For a small number of verbal scales, the strong scalemate is followed with the pronoun ''it'' to make the sentence grammatical. We don't expect this to matter for our purposes.

[6] `https://github.com/hermitdave/FrequencyWords`.

[7] `http://www.opensubtitles.org`.

POS	# unique	Form of original sentence	Form of scalar construction	Example
Adj	120	[NP] is [WEAK]	[NP] is [WEAK], but not [STRONG]	The elephant is big, but not enormous
Adv	12	[NP] is [WEAK] [ADJ]	[NP] is [WEAK] [ADJ], but not [STRONG]	The director is sometimes late, but not always
Verb	16	[NP] [WEAK]-ed	[NP] [WEAK]-ed, but did not [STRONG]	The runner started, but did not finish

Table 3: Scalar construction templates for different parts of speech (for cross-scale variation).

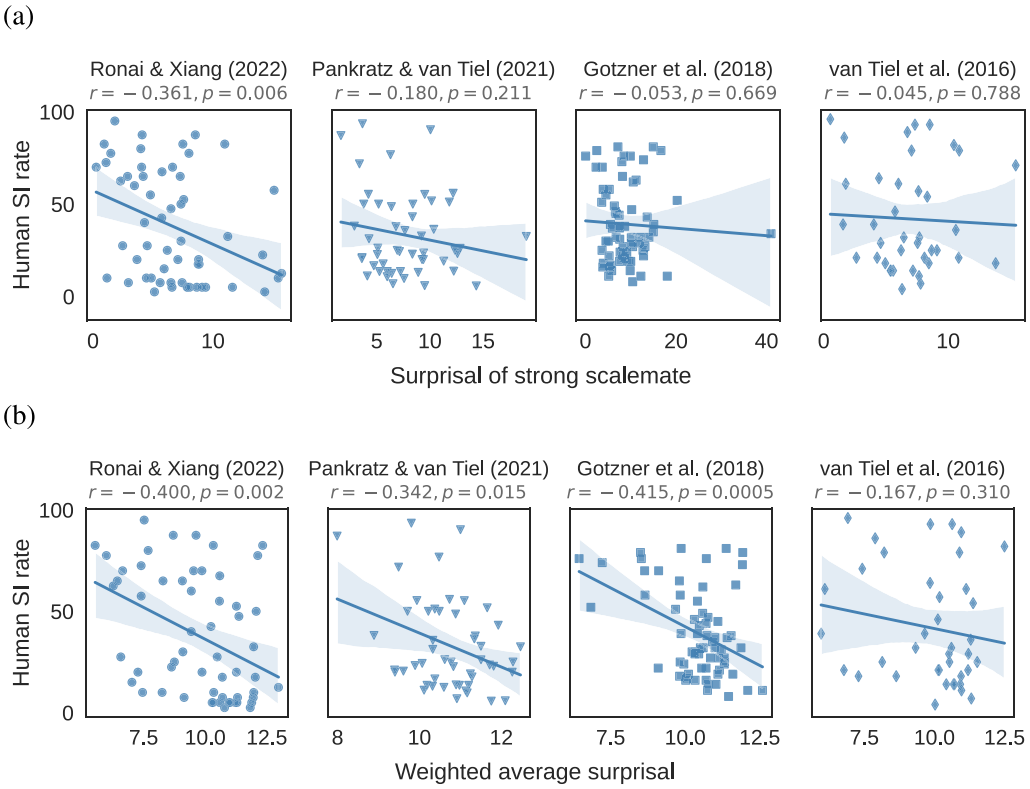

Figure 3: Relationship between human SI rates and GPT-2-derived predictors across scales, for four datasets. Each point represents a single scale. Shaded region denotes 95% CI. (a) SI rate vs. surprisal of strong scalemate in the scalar construction. (b) SI rate vs. weighted average surprisal over the full set of candidate alternatives (Section 5.3).

We excluded words in the POS sets that were not in the frequency corpus, resulting in 3204 adjectives, 1953 adverbs, and 226 verbs. We restricted each POS set to its 1000 highest-frequency words, and performed some manual exclusions (e.g., removing ‘‘do’’ and ‘‘be’’ from the verb set, which are unlikely to form scales with any of the tested items and follow different syntactic rules). This finally resulted in our three alternative sets: 1000 adjectives, 960 adverbs, and 224 verbs.[8]

5.4 Results

5.4.1 String-based Analyses

Figure 3a shows our results for cross-scale variation, under a string-based view of alternatives. We find that surprisal is a significant predictor only for Ronai and Xiang’s dataset (Pearson $\rho = -0.361, p = 0.006$).[9,10]

Model Surprisal vs. Human Completions. For the dataset where we do find a relationship between surprisal and SI rates, we ask whether model surprisals are correlated with human-derived measurements of how ‘‘accessible’’ the strong scalemate is. If model surprisals and human accessibility scores are strongly linked, this would

[8]Most words in the alternative sets occur with low frequency, but we chose to be liberal when including alternatives to ensure broad coverage over potential scalemates.

[9]We repeated this analysis after removing an outlier from Gotzner et al.’s dataset, and again found a lack of relationship between SI rate and surprisal ($\rho = -0.0452, p = 0.719$).

[10]For completeness, we also computed the correlation between SI rates and raw probabilities for both string-based and concept-based analyses (cf. Footnote 4). After excluding outliers, raw probabilities did not achieve stronger Pearson correlations with SI rates than surprisals.

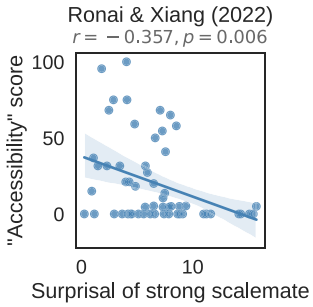

Figure 4: GPT-2-derived surprisal of strong scalemate vs. accessibility rating of strong scalemates (Ronai and Xiang, 2022).

suggest that models and humans are aligned at the level of predictive distributions over alternatives, validating our approach of using language models to approximate human predictions.

To this end, we use data from Ronai and Xiang's Experiment 2, which measured the accessibility of scalemates through a Cloze task. Humans were presented with a short dialogue featuring a sentence with the weak scalemate, as in (3), and then asked to generate a completion of the dialogue in the blank. The ''accessibility'' of the strong scalemate is taken to be the frequency with which it is generated in this paradigm.

Sue: The movie is good. (3)
Mary: So you mean it's not_____.

We find that model surprisals are negatively correlated with accessibility scores (Figure 4; $\rho = -0.357, p = 0.006$), suggesting that our method of estimating expectations over alternatives using artificial language models aligns with direct measurements in humans.

5.4.2 Concept-based Analyses

Turning to a conceptual view of alternatives, Figure 3b shows the relationship between human SI rates and weighted average surprisals (Equation 2). We find a significant negative correlation for all but one of the tested datasets (Ronai and Xiang: $\rho = -0.400, p = 0.002$; Pankratz and van Tiel: $\rho = -0.342, p = 0.015$; Gotzner et al.: $\rho = -0.415, p = 0.0005$; van Tiel et al.: $\rho = -0.167, p = 0.310$), demonstrating that similarity-weighted surprisal captures more variation than raw surprisal (cf. Figure 3a; Section 5.4.1).

We additionally included both (centered) string-based and concept-based surprisal as predictors in a multivariate model, summarized in Table 4 (middle columns). As in the within-scale analysis, for three of the four datasets we find that concept-based surprisal is a stronger predictor than string-based surprisal. With that said, we find only a marginal effect of concept-based surprisal in Ronai and Xiang's data, and no effect of either predictor in van Tiel et al.'s data. However, for Ronai and Xiang's data, this does not mean that there is no value in either predictor—rather, the predictors are too closely correlated to definitively favor one over the other. To demonstrate this, for each dataset we performed an analysis of variance (ANOVA) comparing the full model to a null intercept-only model (Table 4, right columns). We find that for all datasets except that of van Tiel et al., the model with both surprisal predictors explains significantly more variance than the null model. In sum, our results suggest that the expectedness of the strong scalemate can capture significant cross-scale SI variation, but these expectations may operate over groups of semantically similar linguistic forms instead of individual strings.

Qualitative Analysis. As a follow-up analysis, we identified cases where GPT-2 assigns low probability to the tested strong scalemate, but high probability to near synonyms. We analyzed the top 5 alternatives from the full alternative set (Section 5.3) that were assigned highest probability as strong scalemates under GPT-2. Figure 5 shows three examples from Ronai and Xiang's dataset. The title of each subplot shows the scalar construction, with the weak scalemate highlighted in teal and the tested strong scalemate underlined in red. The *y*-axis shows the top 5 candidate scalemates, and the *x*-axis shows the probability assigned by the model. For the weak scalemate *big* (left), GPT-2 assigns highest probability to the alternative *huge*, which conveys similar information to the empirically tested alternative *enormous*. We see a similar pattern for weak scalemate *largely* and alternatives *completely* and *totally* (middle), as well as for weak scalemate *hard* and alternative *impossible* (right). This is consistent with the hypothesis that surprisal of a specific string may not capture surprisal of the underlying concept.

Taken together, these analyses suggest that a concept-based view of alternatives is better

Dataset	Full model			ANOVA	
	Predictor	β	p	F	p
Ronai and Xiang (2022)	String-based surprisal	−1.538	0.215	3.247	0.012
	Concept-based surprisal	−4.503	0.065		
Pankratz and van Tiel (2021)	String-based surprisal	0.460	0.694	3.198	0.050
	Concept-based surprisal	−9.491	0.036		
Gotzner et al. (2018)	String-based surprisal	0.384	0.545	2.751	0.019
	Concept-based surprisal	−8.010	0.0005		
van Tiel et al. (2016)	String-based surprisal	0.293	0.858	1.016	0.422
	Concept-based surprisal	−3.340	0.291		

Table 4: Summary of full regression model (middle columns) and ANOVA comparing full model against intercept-only model (right columns) for each cross-scale variation dataset.

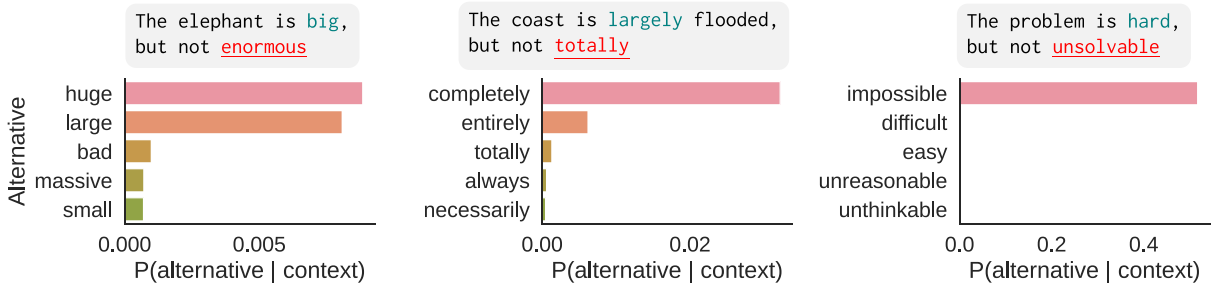

Figure 5: Probability assigned by GPT-2 to top 5 candidate strong alternatives (y-axis) for 3 example weak scalar items: *big*, *largely*, and *hard* (Ronai and Xiang, 2022). The full scalar construction is shown above each subplot, with the original tested strong scalemate underlined in red.

aligned with human inferences than treating alternatives as specific linguistic forms. Testing additional ways of operationalizing concept-based alternatives is a promising direction for future work.

6 Related Work

Prior work has evaluated the ability of computational models to capture scalar inferences. For example, the IMPPRES benchmark (Jeretic et al., 2020) frames SI as a natural language inference problem: The weak scalar expression (e.g., ‘‘Jo ate some of the cake’’) is the premise, and the negated strong scalar expression (e.g., ‘‘Joe didn’t eat all of the cake’’) is the hypothesis. Under this setup, an interpretation consistent with the strictly logical reading would assign a *neutral* relationship between the premise and hypothesis, whereas a pragmatic reading would assign an *entailment* relationship. Models are evaluated based on how often they assign the entailment label across items, which treats SIs as a homogeneous phenomenon and does not capture SI variation.

Another line of work has attempted to predict within-scale SI variation through a supervised approach (Schuster et al., 2020; Li et al., 2021). This approach takes a sentence with a weak scalar item, and attempts to directly predict the human SI strength through a prediction head on top of a sentence encoder. This differs from our approach in that it requires training directly on the SI-rate-prediction task, whereas we probe the predictive distribution that emerges from language modeling with no task-specific representations. This allows us to compare model probability distributions to the expectations deployed by humans during pragmatic inferences, building upon a literature linking language models to predictive processing (e.g., Frank and Bod, 2011; Smith and Levy, 2013; Wilcox et al., 2020; Merkx and Frank, 2021).

There have also been several studies extracting scalar orderings from corpora or language model representations. For example, de Marneffe et al. (2010) use distributional information from a web corpus to ground the meanings of adjectives for an indirect question answering task. Similarly, Shivade et al. (2015) use scalar constructions like ''*X, but not Y*'' to identify scales from a corpus of biomedical texts. Others have found that adjectival scale orderings can be derived from static word embeddings (Kim and de Marneffe, 2013) and contextualized word representations (Garí Soler and Apidianaki, 2020; Garí Soler and Apidianaki, 2021).

7 Discussion

We tested a shared mechanism explaining variation in SI rates across scales and within ⟨*some*, *all*⟩, based on the hypothesis that humans maintain context-driven expectations about unspoken alternatives (Degen and Tanenhaus, 2015, 2016). We operationalized this in two ways using neural language models: the expectedness of a linguistic alternative as a scalemate (string-based surprisal), and the expectedness of a conceptual alternative (weighted average surprisal). We found that for both within-scale and cross-scale variation, expectedness captures human SI rates. Crucially, however, expectedness of the strong scalemate is a robust predictor of cross-scale variation only under a conceptual view of alternatives (Buccola et al., 2021). Our results support the idea that the strength of pragmatic inferences depends on the availability of alternatives, which depends on in-context predictability.

One open question is the source of variability across the tested human behavioral datasets—in particular, the lack of surprisal effect for van Tiel et al.'s data (Section 5.4). While we cannot be certain about why the results vary, we identified a few differences that might affect data quality across datasets (see Table 1). van Tiel et al.'s study has the smallest number of participants (28), smallest number of ratings per scale (10), and smallest number of scales (39). In addition, their experiments presented multiple sentence contexts per scale, whereas the other experiments only presented one sentence per scale. Other experimental factors, such as participant recruitment and exclusion criteria, may have also contributed to differences in data reliability.

7.1 How Do Listeners Restrict the Alternatives?

We now return to the issue raised in Footnote 2: what information do listeners use to form expectations about alternatives? To illustrate potential hypotheses, consider the item ''The soup is warm/hot'' from van Tiel et al.'s experimental materials. In our framework described in Section 3.1, `[CONTEXT]` = ''The soup is'', `[WEAK]` = ''warm'', and `[STRONG]` = ''hot''. One hypothesis is that listeners form expectations over relevant scalar expressions given `[CONTEXT]` alone. On this view, expectations over strong scalemates could be measured by computing the probability of `[STRONG]` in the template `[CONTEXT][STRONG]`; i.e., ''The soup is `[STRONG]`''. In contrast, in this paper we test expectations of `[STRONG]` in the template ''The soup is warm, but not `[STRONG]`'', which instantiates an alternate theoretical position: that listeners use not only the context, but also `[WEAK]` as information for forming expectations over alternatives.

We adopt this view for several reasons. First, it could be the case that the context does not provide enough information for the listener to narrow down alternatives. Returning to the running example, ''The soup is'' could be followed by many continuations, some potentially relating to the taste or size of the soup in addition to its temperature. Taking the weak scalar term ''warm'' into account allows the listener to restrict the relevant alternatives to a smaller, more tractable set, which presents an algorithmic solution to the computationally challenging inference problem. However, the underinformativity of the context may be a problem unique to the simple stimuli used in the behavioral experiments. It is plausible that listeners could sufficiently restrict alternative sets given more naturalistic contexts, which likely provide more cues to the Question Under Discussion (Roberts, 2012).

In addition, there could be cues from `[WEAK]` that provide information about likely alternatives, independent of the context. For example, listeners might prefer strong scalemates that match `[WEAK]` in register or formality, or in shared phonological features. This motivates why we chose template (1) to measure expectations over alternatives, instead of `[CONTEXT][STRONG]`. However, the extent to which listeners tune their

predictions based on [WEAK] above and beyond the context remains an open empirical question.

7.2 From Alternatives to Inference

Conceptually, computing an SI involves two steps: (1) determining the suitable alternatives, and (2) ruling out the meaning of alternatives to arrive at a strengthened interpretation of the weak scalar term. Our results primarily shed light on the first step, providing evidence that expectations play a role in determining alternatives, and that alternatives are likely based on meanings in addition to linguistic forms.

When considering the higher-level reasoning process, many factors beyond alternatives play a causal role in SI. One view is that humans use alternatives in a cooperative reasoning process, such as that formalized by the Rational Speech Act framework (RSA; Frank and Goodman, 2012; Goodman and Frank, 2016). In an RSA model, a pragmatic listener $L_1(m \mid u)$ uses a speaker's utterance u to update their prior beliefs $P(m)$ over which meaning m the speaker is trying to convey. The listener does this by computing the likelihood of a pragmatic speaker S_1 producing u given each potential meaning. The pragmatic S_1 speaker corresponds to the utility U of the utterance u to convey m, relative to the utility of the alternative utterances in the set of alternatives $\mathcal{A}$:

$$L_1(m \mid u) = \frac{S_1(u \mid m)P(m)}{\sum_{m'} S_1(u \mid m')P(m')} \quad (4)$$

$$S_1(u \mid m) = \frac{U(u, m)}{\sum_{u' \in \mathcal{A}} U(u', m)} \quad (5)$$

Our findings appear compatible with RSA: Listeners reason about a speaker that normalizes over alternatives. However, it remains an open question how variable expectations over alternatives should be operationalized in an RSA model. One option, as recently proposed by Zhang et al. (2023), is that the pragmatic speaker is conditioned on the alternative set $\mathcal{A}$. The pragmatic listener has beliefs over different sets of $\mathcal{A}$ and marginalizes over these beliefs when drawing an inference:

$$L_1(m \mid u) = \sum_{\mathcal{A}} P(\mathcal{A}) \frac{S_1(u \mid m, \mathcal{A})P(m)}{\sum_{m'} S_1(u \mid m', \mathcal{A})P(m')} \quad (6)$$

Another possibility is that the variable expectations are not inputs to the model, but instead fall out of reasoning about how likely speakers are to use the weaker versus stronger terms, given variable contextual priors over meanings and questions under discussion (see, e.g., Goodman and Lassiter, 2015; Qing et al., 2016). We leave a detailed exploration of such a model to future work.

The Role of Priors. Pragmatic inferences are influenced by the prior probabilities of the world states compatible with the weak and strong meanings (Degen et al., 2015; Sikos et al., 2021). For example, consider the scale ⟨*start*, *finish*⟩. If a human were asked ''The movie started at 2:30. Would you conclude that the movie did not finish at 2:30?'', they would likely answer *Yes*. This *Yes* response would count as an SI under the experimental paradigm, but does not reflect pragmatic reasoning over scalar alternatives: It is simply implausible for a movie to start and finish at the same time, given our knowledge of the world.[11]

These priors have an important connection to our analyses. As outlined in Section 3.1, we approximate the expectedness of a strong scalemate by measuring the expectedness of its linguistic form. This approach can be seen as reflecting an implicit assumption that the more likely a certain meaning is, the more likely it is to be expressed linguistically. This is likely to be wrong in certain cases—for example, if a certain meaning is so likely that it is obvious without being said, then speakers may avoid the effort of explicitly producing the linguistic expression (and thus, the linguistic expression would have low probability). This could potentially be the case for relatively common SIs. For example, a speaker might be able to get away with only saying *some* and expecting a listener to recover the meaning *some but not all*.

We believe our estimation method may minimize this issue, as we measure expectations conditioned on an explicit scalar contrast with the weak scalemate (i.e., ''[WEAK], but not''). Thus, our approach can be seen as approximating listeners' expectations, given that the speaker has *already chosen* to produce a scalar contrast. Nevertheless, a complete account of scalar inferences will need to account for the influence of the

[11] This example is due to Lassiter (2022).

prior probabilities over world states, which may explain some of the variance not captured by our predictors.

7.3 Implications for NLP

While the main role of language models in our analyses was to systematically test a cognitive theory, we believe this work also has implications for NLP evaluation. A growing body of work uses controlled assessments to evaluate the linguistic knowledge of NLP models. Many studies test whether models exhibit a categorical pattern of behavior that reflects a particular linguistic generalization. For example, in syntactic evaluations, a model is successful if it satisfies certain inequality relationships between grammatical and ungrammatical sentences (e.g., Linzen et al., 2016; Futrell et al., 2019; Hu et al., 2020). SI (and other types of implicatures) have largely been treated the same way (see Section 6).

In contrast, we do not evaluate whether language models exhibit a categorical pattern of behavior (*''Do models interpret SIs pragmatically?''*). Instead, based on the empirical evidence for scalar variation, we test whether models capture systematic variability in human inferences (*''Are models sensitive to the factors that modulate human pragmatic inferences?''*). We urge other NLP researchers to consider variability in human behaviors instead of relying on categorical generalizations (see also Pavlick and Kwiatkowski, 2019; Jiang and Marneffe, 2022; Baan et al., 2022; Webson et al., 2023). Through this approach, we can build models that capture the rich variability of human language, and use these models to refine our theories about the human mind.

Acknowledgments

We thank the anonymous reviewers and the action editor for their insightful feedback. J. H. is supported by an NSF Graduate Research Fellowship (#1745302) and an NSF Doctoral Dissertation Research Improvement Grant (BCS-2116918). S.S. is supported by the NSF under grant #2030859 to the Computing Research Association for the CIFellows Project and the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 948878). J. H. and R. L. also gratefully acknowledge support from the Simons Center for the Social Brain at MIT. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation nor the Computing Research Association.

References

Joris Baan, Wilker Aziz, Barbara Plank, and Raquel Fernandez. 2022. Stop measuring calibration when humans disagree. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1892–1915, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://aclanthology.org/2022.emnlp-main.124/

Andrea Beltrama and Ming Xiang. 2013. Is 'good' better than 'excellent'? An experimental investigation on scalar implicatures and gradable adjectives. *Proceedings of Sinn und Bedeutung*, 17.

Richard Breheny, Nathan Klinedinst, Jacopo Romoli, and Yasutada Sudo. 2018. The symmetry problem: Current theories and prospects. *Natural Language Semantics*, 26(2):85–110. https://doi.org/10.1007/s11050-017-9141-z

Brian Buccola, Manuel Križ, and Emmanuel Chemla. 2021. Conceptual alternatives: Competition in language and beyond. *Linguistics and Philosophy*. https://doi.org/10.1007/s10988-021-09327-w

Joan L. Bybee and Clay Beckner. 2015. Usage-based theory. In Bernd Heine and Heiko Narrog, editors, *The Oxford Handbook of Linguistic Analysis*. Oxford University Press.

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. https://doi.org/10.3765/sp.8.11

Judith Degen and Michael K. Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive Science*, 39(4):667–710. https://doi.org/10.1111/cogs.12171, PubMed: 25265993

Judith Degen and Michael K. Tanenhaus. 2016. Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, 40(1):172–201. https://doi.org/10.1111/cogs.12227, PubMed: 25807866

Judith Degen, Michael Henry Tessler, and Noah D. Goodman. 2015. Wonky worlds: Listeners revise world knowledge when utterances are odd. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

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. https://doi.org/10.1163/187730909X12538045489854

Sarah F. V. Eiteljoerge, Nausicaa Pouscoulous, and Elena V. M. Lieven. 2018. Some pieces are missing: Implicature production in Children. *Frontiers in Psychology*, 9:1928. https://doi.org/10.3389/fpsyg.2018.01928, PubMed: 30405468

Danny Fox and Roni Katzir. 2011. On the characterization of alternatives. *Natural Language Semantics*, 19(1):87–107. https://doi.org/10.1007/s11050-010-9065-3

Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998–998. https://doi.org/10.1126/science.1218633, PubMed: 22628647

Stefan L. Frank and Rens Bod. 2011. Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6):829–834. https://doi.org/10.1177/0956797611409589, PubMed: 21586764

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1004

Aina Garí Soler and Marianna Apidianaki. 2020. BERT knows Punta Cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualised representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7371–7385, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.598

Aina Garí Soler and Marianna Apidianaki. 2021. Scalar adjective identification and multilingual ranking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4653–4660, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.370

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.10

Gerald Gazdar. 1979. *Pragmatics: Implicature, Presupposition, and Logical Form*. Academic Press, New York.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. 1992. Switchboard: A telephone speech corpus for research and development. In *International Conferenceon Acoustics, Speech and Signal Processing*, pages 517–520. https://doi.org/10.1109/ICASSP.1992.225858

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829. https://doi.org/10.1016/j.tics.2016.08.005, PubMed: 27692852

Noah D. Goodman and Daniel Lassiter. 2015. Probabilistic semantics and pragmatics: Uncertainty in language and thought. In *The Handbook of Contemporary Semantic Theory*, pages 655–686. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118882139.ch21

Nicole Gotzner, Stephanie Solt, and Anton Benz. 2018. Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology*, 9:1659. https://doi.org/10.3389/fpsyg.2018.01659

Herbert P. Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Speech Acts*, 3, pages 41–58. Academic Press. https://doi.org/10.1163/9789004368811_003

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*. https://doi.org/10.3115/992133.992154

Laurence R. Horn. 1989. *A Natural History of Negation*. Chicago University Press.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.158

Paloma Jeretic, Alex Warstadt, Suvrat Bhooshan, and Adina Williams. 2020. Are natural language inference models IMPPRESsive? Learning IMPlicature and PRESupposition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8690–8705. Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.768

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374. https://doi.org/10.1162/tacl_a_00523

Roni Katzir. 2007. Structurally-defined alternatives. *Linguistics and Philosophy*, 30(6):669–690. https://doi.org/10.1007/s10988-008-9029-y

Joo-Kyung Kim and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1625–1630. Seattle, Washington, USA. Association for Computational Linguistics. https://aclanthology.org/D13-1169

Anthony Kroch. 1972. Lexical and inferred meanings for some time adverbs. *Quarterly Progress Reports of the Research Laboratory of Electronics*, 104:260–267.

Daniel Lassiter. 2022. How not to identify a scalar implicature (The importance of priors). Presentation at Cognitive Semantic and Quantities Workshop, University of Amsterdam.

Stephen Levinson. 2000. *Presumptive Meaning: The Theory of Generalized Conversational Implicature*, MIT Press. https://doi.org/10.7551/mitpress/5526.001.0001

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. https://doi.org/10.7275/xr01-a852

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535. https://doi.org/10.1162/tacl_a_00115

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. https://aclanthology.org/L16-1147/

Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics*, pages 63–70, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics. https://doi.org/10.3115/1118108.1118117

Marie-Catherine de Marneffe, Christopher D. Manning, and Christopher Potts. 2010. ''Was it good? It was provocative.'' Learning the meaning of scalar adjectives. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 167–176, Uppsala, Sweden. Association for Computational Linguistics. https://aclanthology.org/P10-1018/

David Marr. 1982. *Vision: A Computational Approach*. Freeman & Co., San Francisco.

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. https://doi.org/10.1162/tacl_a_00227

Danny Merkx and Stefan L. Frank. 2021. Human sentence processing: Recurrence or attention? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 12–22. Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.cmcl-1.2

G. A. Miller. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41. https://doi.org/10.1145/219717.219748

Emiel van Miltenburg. 2015. Detecting and ordering adjectival scalemates. In *Proceedings of MAPLEX*, Yamagata, Japan. https://arxiv.org/abs/1504.08102

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. https://doi.org/10.1017/langcog.2021.13

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694. https://doi.org/10.1162/tacl_a_00293

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543. Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Ciyang Qing, Noah D. Goodman, and Daniel Lassiter. 2016. A rational speech-act model of projective content. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69. https://doi.org/10.3765/sp.5.6

Eszter Ronai and Ming Xiang. 2021. Exploring the connection between question under discussion and scalar diversity. In *Proceedings of the Linguistic Society of America*, volume 6, pages 649–662. https://doi.org/10.3765/plsa.v6i1.5001

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

Mats E. Rooth. 1985. *Association with Focus*. PhD thesis, University of Massachusetts.

Sebastian Schuster, Yuxing Chen, and Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5387–5403, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.479

Chaitanya Shivade, Marie-Catherine de Marneffe, Eric Fosler-Lussier, and Albert M. Lai. 2015. Corpus-based discovery of semantic intensity scales. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–493. Denver, Colorado. Association for Computational Linguistics. https://doi.org/10.3115/v1/N15-1051

Les Sikos, Noortje J. Venhuizen, Heiner Drenhaus, and Matthew W. Crocker. 2021. Reevaluating pragmatic reasoning in language games. *PLOS ONE*, 16(3):e0248388. https://doi.org/10.1371/journal.pone.0248388, PubMed: 33730097

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319. https://doi.org/10.1016/j.cognition.2013.02.013, PubMed: 23747651

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. https://doi.org/10.3389/fpsyg.2018.02092, PubMed: 30443233

Bob van Tiel, Emiel van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of Semantics*, 33(1):137–175. https://doi.org/10.1093/jos/ffu017

Michael Tomasello. 2003. *Constructing A Language: A Usage-based Theory of Language Acquisition*, Harvard University Press.

Albert Webson, Alyssa Marie Loo, Qinan Yu, and Ellie Pavlick. 2023. Are language models worse than humans at following prompts? It's complicated. https://doi.org/10.48550/arXiv.2301.07085

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 42nd Annual Meeting of the Cognitive Science Society*. https://arxiv.org/abs/2006.01912

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. https://doi.org/10.18653/v1/2020.emnlp-demos.6

Zheng Zhang, Leon Bergen, Alexander Paunov, Rachel Ryskin, and Edward Gibson. 2023. Scalar implicature is sensitive to contextual alternatives. *Cognitive Science*, 47(2). https://doi.org/10.1111/cogs.13238, PubMed: 36739521