Scalar diversity

Bob van Tiel
Emiel van Miltenburg
Natalia Zevakhina
Bart Geurts

Abstract

We present experimental evidence showing that there is considerable variation between the rates at which sentences containing scalar expressions give rise to upper-bounded construals. In two experiments, we found that the frequency of scalar inferences was much higher for quantifying and modal expressions than for adjectives and verbs. However, there was no dichotomous distinction between these categories, since several adjectives and verbs gave rise to elevated levels of scalar inferences, too. We investigated four factors that might be responsible for the variation between scalar expressions, namely focus, word frequency, strength of association, and semantic distance. It was found that only semantic distance, which is a measure of the perceived difference in strength between a sentence and its alternative, had a significant effect on the rates of scalar inferences. We propose that our results are best explained as being shaped by general considerations of relevance, rather than in terms of discourse goals or questions under discussion.

Keywords: scalar implicature, quantity implicature, experimental pragmatics

Introduction

A speaker who says (1) usually implies that she did not eat all of the cookies. The scalar expression 'some', whose logical meaning is just 'at least some', receives an upper-bounded interpretation and thus comes to exclude 'all'.

(1) I ate some of the cookies.

To explain this *scalar inference*, it is often assumed that scalar expressions evoke scales whose members are ordered in terms of informativeness. For instance, 'some' evokes the scale ⟨some, all⟩, where 'all' is more informative than 'some'. A speaker who uses a less than maximally informative scalar expression implies, at least in

some situations, that she does not believe that one of the more informative scalar expressions would have been appropriate. Table 1 lists some of the scales that have been proposed in the literature.[1]

Category	**Examples**	
Verbs	⟨might, must⟩	⟨believe, know⟩
	⟨like, love⟩	⟨start, finish⟩
Adjectives	⟨good, excellent⟩	⟨intelligent, brilliant⟩
	⟨difficult, impossible⟩	⟨cheap, free⟩
Adverbs	⟨sometimes, always⟩	⟨possibly, necessarily⟩
Nouns	⟨mammal, dog⟩	⟨lecturer, professor⟩
Connectives	⟨or, and⟩	⟨if, if and only if⟩
Determiners	⟨some, all⟩	⟨few, none⟩

Table 1: Sample scales for various lexical categories.

The debate about scalar inferences has, for the most part, centered on the question how these inferences come about. The traditional answer is that scalar inferences are a variety of conversational implicature (cf. Horn 1972). Someone who hears (1) first interprets 'some' as meaning 'at least some'. She then observes that the speaker could have been more informative by saying that she ate all of the cookies. Why didn't she do so? Presumably because she did not eat all of the cookies. However, several authors have proposed alternatives to this account. Levinson (2000), for example, stipulates that scalar terms are ambiguous between a logical and an upper-bounded interpretation; so 'some' is ambiguous between meaning 'at least some' and 'some but not all'. Hearers arrive at the upper-bounded interpretation by default, which can subsequently be cancelled under special circumstances.

A fair number of experiments have been done to compare the predictions of various theories. One striking feature of these experiments is that, for the most part, they are confined to just two scalar expressions, namely 'some' and 'or'. To illustrate, Table 2 provides an overview of the scalar expressions that have been used in a representative sample of the research on the development and processing of scalar inferences. A comparison with Table 1 makes it clear that several classes of scalar expressions, most notably nouns, adjectives and adverbs, have been consistently overlooked. Even within the classes that have been investigated, the variety of scalar expressions is limited. Apparently, the tacit assumption underlying these

1. This overview does not include some expressions whose status is controversial, in particular, articles (e.g., Horn 2007), numerals (e.g., Spector 2012), and the plural morpheme (e.g., Zweig 2009). This is not to imply, however, that all the scales listed here are uncontroversial.

Scale	Sources	
⟨some, all⟩	Noveck (2001)	Noveck & Posada (2003)
	Papafragou & Musolino (2003)	Bott & Noveck (2004)
	Feeney et al. (2004)	Guasti et al. (2005)
	Breheny et al. (2006)	De Neys & Schaeken (2007)
	Pouscoulous et al. (2007)	Huang & Snedeker (2009)
	Grodner et al. (2010)	Barner et al. (2011)
	Bott et al. (2012)	
⟨or, and⟩	Noveck et al. (2002)	Storto & Tanenhaus (2005)
	Breheny et al. (2006)	Chevallier et al. (2008)
	Pijnacker et al. (2009)	Zondervan (2010)
⟨might, must⟩	Noveck (2001)	
⟨start, finish⟩	Papafragou & Musolino (2003)	

Table 2: Scalar expressions used in a representative sample of experiments on the development and processing of scalar inferences.

experiments is that the scalar expressions in Table 2, and especially 'some' and 'or', are representative for the entire family of scalar expressions. Until recently, the *uniformity assumption*, as we will call it, had not been questioned, but it was put to the test by Doran, Baker, McNabb, Larson, and Ward (2009), following up on a study by the same group (Larson et al. 2009). Doran et al.'s findings suggest that there is significant variability between the rates at which scalar terms of different lexical categories give rise to upper-bounded inferences. However, as we will argue in the following, Doran et al.'s experimental design precludes a straightforward interpretation of their data. We therefore undertook a study based on a simpler design, which also provided us with finer-grained results than previous studies. Furthermore, we investigated several candidate explanations for the variability we observed.

1. Extant evidence for diversity

Preliminary evidence against the uniformity assumption is provided by Geurts's (2010, 98-99) survey of ten experiments employing the verification paradigm. In these experiments, participants had to decide whether target sentences were true or false in a given state of affairs. For example, Bott and Noveck's (2004, experiment 3) participants rejected statements like (2a,b) 59% of the time:

(2) a. Some parrots are birds.
b. Some dogs are mammals.

Supposing that this figure is a fair estimate of the rate at which scalar inferences were derived (which may not be entirely obvious), this means that, for sentences like (2a,b), Bott and Noveck's participants derived scalar inferences 59% of the time. The other experiments included in Geurts's survey used similar designs, though there were important differences between studies with respect to the way states of affairs were presented: whereas Bott and Noveck relied on participants' world knowledge, many other studies used pictures instead. For our current purposes, the main point transpiring from Geurts's survey is that, across the collated experiments, the mean rate of scalar inferences for disjunction was clearly lower than for 'some': 35% against 56.5%. This suggests rather strongly that scalar inference rates are higher for 'some' than for 'or'.

Doran et al. (2009) were the first to test the uniformity assumption in an integrated experimental design.[2] They presented participants with stories like the following:

(3) Irene: How much cake did Gus eat at his sister's birthday party?
Sam: He ate most of it.
FACT: By himself, Gus ate his sister's entire birthday cake.

(4) Irene: How attractive is Kate?
Sam: She's pretty.
FACT: Kate was voted 'World's Most Beautiful Woman' this year.

Participants had to decide whether Sam's answers were true or false, as seen from the perspective of one 'Literal Lucy', a literal-minded character supposedly concerned with truth-conditional meaning only. Literal Lucy was introduced on the premiss that if Sam's statement was deemed to be false, from Literal Lucy's perspective, then the scalar inference associated with the scalar term in his answer must have been part of its truth-conditional meaning. Hence, strictly speaking, participants were not expected to proffer their own judgments; rather, they had to decide how Literal Lucy would interpret Sam's statements.

One further complication introduced by Doran et al. was that, in addition to the condition illustrated in (3) and (4), there were two further conditions: one in which Irene's question contained a scalar term that was stronger than the one used by

2. As we mentioned in our introduction, Doran et al.'s study followed up on a similar one by the same group (Larson et al. 2009). The methodology of the two studies was the same, and their results were similar. We will confine our attention to Doran et al.'s study, because its objective was more closely related to ours. Our description of Doran et al.'s experimental design will simplify matters somewhat, leaving out parts that are irrelevant here.

Sam in his answer, as in (5a) and (6a), and one in which Irene's question, in effect, offered Sam three scalar expressions to choose from, as in (5b) and (6b):

(5) a. Did Gus eat all of his sister's birthday cake?
    b. Did Gus eat some, most, or all of his sister's birthday cake?

(6) a. Is Kate gorgeous?
    b. Is Kate average-looking, pretty, or gorgeous?

In the following, we will use the terms 'neutral' and '(one- or two-way) contrastive' to label these conditions: (3) and (4) count as neutral, (5a) and (6a) are one-way contrastive, and (5b) and (6b) are two-way contrastive.

Doran et al.'s first main finding was that, whereas quantified statements were rejected about half of the time, for sentences with adjectives, the rejection rate was about half that (no exact figures given). Doran et al. interpret this result as showing that scalar inferences associated with quantifiers were incorporated in the truth-conditional content of Sam's answers twice as often as adjectival scalar inferences. Secondly, Doran et al. found that only adjectival items were affected by the difference between neutral and contrastive conditions: within the adjectival category, the two-way contrastive items elicited significantly more 'false' responses than the neutral and the one-way contrastive ones, although this effect was quite modest (only slightly more than 10%); otherwise, the neutral/contrastive distinction was inert. It appears therefore that, overall, the neutral/contrastive distinction had very little effect on the interpretation of Sam's answer.

Both Doran et al.'s experiment and the proposed interpretation of its results are open to criticism. We will discuss three issues here. First, Doran et al. wanted their participants to judge Sam's answers on behalf of a third party, 'Literal Lucy'. They introduced this extra layer of complexity on the assumption that it provided a means of distinguishing between truth-conditional and non-truth-conditional content. However, as discussed by Geurts (2010, 7-9), truth-conditional content is a theoretical notion which does not necessarily align with naive hearers' intuitions about truth. A more pressing worry is that Literal Lucy introduced a task requirement whose effect is hard to gauge (it is not at all clear, for example, that Doran et al.'s participants managed to consistently adopt Literal Lucy's perspective), and most importantly: to the extent that the introduction of Literal Lucy had its intended effect, Doran et al.'s participants were not giving their own judgements on Sam's answers; rather, they were indicating how they thought Literal Lucy would interpret them. Put otherwise, strictly speaking it is contestable that Doran et al.'s data tell us anything about how their participants themselves interpreted the sentences they were asked to judge.

Secondly, Doran et al.'s experiment employed a verification task for gauging the frequency of scalar inferences, but it is unique in that it presented the relevant facts

by way of verbal description. This mode of presentation is bound to be non-uniform. For example, according to our intuitions, Doran et al.'s fact descriptions for adjectival items imply less strongly that Sam's utterance is too weak than the descriptions they presented with quantificational items; perhaps because the descriptions for quantificational items always contained an expression that was synonymous with the stronger alternative, like 'entire' in (3). This intuition may be debatable, but the fact remains that verbal descriptions are harder to standardise than pictures, for example. Thirdly, Doran et al. adopted a rather coarse-grained categorization of experimental items, lumping together quantifying expressions with measure phrases and modal adverbs, for example. The fact that they found a dichotomous distinction between quantifying and adjectival expressions may have been due to this, and it is quite possible that a finer-grained analysis would have produced results that speak against such a dichotomy. The bottom line is that, although Doran et al.'s findings provide prima facie evidence against the uniformity assumption, there are good reasons for going over the same ground with a cleaner and simpler experimental design and a finer-grained analysis—which is what we did.

2. New evidence for diversity

Since the verification paradigm is ill-suited for testing a sufficiently rich variety of scalar expressions, we decided to adopt an inference paradigm, which has been widely used in the psychology of reasoning, and has occasionally been used in experimental studies on scalar inference. It has been shown that the inference paradigm yields higher rates of scalar inferences than the verification paradigm (Geurts & Pouscoulous 2009), but since we were primarily interested in relative frequencies of scalar inferences, that was no cause for concern. Still, as we will presently see, our data forced us to have a closer look at the inference paradigm.

2.1. Experiment 1

Participants

We posted surveys for 25 participants on Amazon's Mechanical Turk (mean age: 38; range: 21–63; 14 females).[3] Only workers with an IP address from the United States were eligible for participation. In addition, these workers were asked to indicate their native language. Payment was not contingent on their response to this question. It turned out that all of our participants in this experiment were native

3. Mechanical Turk is a website where workers perform so-called 'Human Intelligence Tasks' (HITs) for financial compensation. It has been shown that the quality of data gathered through Mechanical Turk equals that of laboratory data (Schnoebelen & Kuperman 2010, Buhrmester, Kwang, & Gosling 2011, Sprouse 2011).

speakers of English. Participants received $0.60 for their participation, which took approximately 10 minutes.

Materials and procedure

Figure 1 shows an example of a critical item (the full list of materials is given in the appendix). In each trial, a character named John or Mary made a statement containing a scalar expression, which always occurred in predicate position, and participants had to decide whether or not this implied that, according to the speaker, the statement would have been false if that expression had been replaced with a stronger scale member. The statements were kept as bland as possible, so that participants would not be guided by expectations based on their world knowledge. This was done mainly by using pronouns instead of complex noun phrases, but also by using generic predicates like 'go inside' and 'do that'. (Experiment 2, which is reported in the next section, replicated the current experiment with more informative sentences.) Pronouns were never congruent with the speaker's gender in order to prevent them from being interpreted as referring to the speaker.

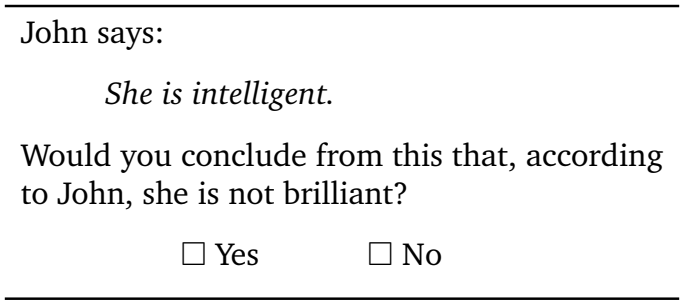

Figure 1: Sample item used in Experiment 1.

Materials comprised a selection of scales consisting of quantifiers (3), modal expressions (4), adjectives (30), and main verbs (6). A complete list is given in Table 3. The selection of scales consisting of adjectives can be further subdivided depending on whether the stronger scale member denotes an end point, in which case the scale is 'closed'; otherwise it is 'open'. For instance, ⟨difficult, impossible⟩ is closed, because 'impossible' denotes an end point; ⟨ugly, hideous⟩ is open, because 'hideous' does not denote an end point. We will see that the distinction between open and closed scales is relevant to the derivation of scalar inferences.[4]

4. It might be questioned that all the expressions with an end-point alternative are really scalars. For instance, it might be argued that if something is extinct it is not rare. Several considerations speak against this possibility. Firstly, it is perfectly felicitous to describe something as being 'rare, if not

Though we are primarily interested in the distinction between open and closed adjective scales, it should be noted that the distinction applies to other categories, too. It is clear, for instance, that 'all' denotes the end point of the ⟨some, all⟩ scale. Thus understood, about half of our experimental items involved open scales, while the other half involved closed scales. However, no attempt was made to balance open and closed items in the non-adjectival categories.

Our selection of scalar expressions was guided in part by examples discussed in the literature (e.g., Horn 1972, Hirschberg 1991, Doran et al. 2009). However, adjectival scales, which were used in 70% of the experimental items, were selected by searching the internet and several corpora (the British National Corpus, the Corpus of Contemporary American English, and the Open American National Corpus) for constructions of the form 'X if not Y', 'X or even Y', and 'not just X but Y', which yielded a large number of candidate scales. The final selection included, besides ten closed scales, ten scales whose weaker term occurred more frequently than the stronger term, based on word counts in the Corpus of Contemporary American English (Davies 2008), and ten scales for which the opposite was true; we did this because we wanted to test the hypothesis that frequency has an effect on the rate at which a scalar inference is derived (Section 4).

Randomised lists were created for each participant, varying the order of the items. Seven control items were included, which involved statements that either entailed (e.g., an inference from 'wide' to 'not narrow') or were completely unrelated to (e.g., an inference from 'sleepy' to 'not rich') the critical inference (see appendix). In a pilot experiment we gauged whether the number of control items had an effect on the results of the inference task. We presented 50 participants (mean age: 35; range: 18–67; 30 females) on Mechanical Turk with 10 of the target items included in Experiment 1 alongside 32 control items. In 16 of the control trials, the target inference was clearly valid; in the remaining 16 controls, it was clearly not valid.

---

extinct'; a Google search yields several hundred occurrences, like the following:

(i) a. Originality is now rare, if not extinct.
b. The black-striped wallaby is now rare, if not extinct, on Kawau.

Secondly, these items are structurally the same as bona fide scales like ⟨some, all⟩ and ⟨possible, certain⟩. In these cases, the stronger scalar term denotes an end point as well, and in these cases, too, many people who are not familiar with the notion of scalar inferences have the intuition that, for example, if I ate all of the cookies, I did not eat some of them. Thirdly, we believe that inference patterns like the following are valid:

(ii) X is rare.
Y occurs less frequently than X.
Y is rare.

Such inference patterns are predicted to be invalid if 'rare' excludes 'extinct' by entailment.

Scale	Category	SI		Cloze		LSA	Dist	Freq
		+N	–N	+N	–N			
⟨adequate, good⟩	Adjective (o)	29	32	33	57	.27	3.52	-1.52
⟨allowed, obligatory⟩	Modal	67	82	20	47	.02	5.35	-0.85
⟨attractive, stunning⟩	Adjective (o)	8	21	53	72	.07	5.78	0.37
⟨believe, know⟩	Verb	21	61	67	67	.46	5.04	-0.70
⟨big, enormous⟩	Adjective (o)	17	21	83	37	.21	5.43	1.13
⟨cheap, free⟩	Adjective (c)	100	93	0	0	.19	5.52	-0.66
⟨content, happy⟩	Adjective (o)	4	4	87	50	.13	4.52	-0.85
⟨cool, cold⟩	Adjective (o)	33	46	23	40	.61	4.30	-0.21
⟨dark, black⟩	Adjective (c)	4	29	30	27	.40	4.04	-0.49
⟨difficult, impossible⟩	Adjective (c)	79	96	13	10	.60	6.22	0.46
⟨dislike, loathe⟩	Verb	29	18	93	90	.16	5.87	0.46
⟨few, none⟩	Quantifier	75	54	20	30	.47	5.35	0.75
⟨funny, hilarious⟩	Adjective (o)	4	29	50	33	.07	5.04	1.17
⟨good, excellent⟩	Adjective (o)	37	32	60	57	.46	5.48	1.34
⟨good, perfect⟩	Adjective (c)	46	39	60	23	.42	6.09	1.00
⟨hard, unsolvable⟩	Adjective (c)	71	71	10	10	.08	5.26	2.87
⟨hungry, starving⟩	Adjective (o)	33	25	63	40	.52	5.74	0.71
⟨intelligent, brilliant⟩	Adjective (o)	8	7	17	3	.27	4.74	-0.12
⟨like, love⟩	Verb	50	25	80	57	.37	5.74	0.23
⟨low, depleted⟩	Adjective (c)	71	79	23	60	.16	4.87	2.29
⟨may, have to⟩	Modal	75	71	83	80	.64	5.26	-1.22
⟨may, will⟩	Modal	87	89	83	80	.51	5.41	0.68
⟨memorable, unforgettable⟩	Adjective (c)	50	54	23	60	.29	4.83	0.56
⟨old, ancient⟩	Adjective (o)	17	36	50	33	.24	5.39	1.08
⟨palatable, delicious⟩	Adjective (o)	58	61	67	47	.32	5.52	-0.89
⟨participate, win⟩	Verb	21	18	7	37	.21	6.35	-0.62
⟨possible, certain⟩	Modal	92	93	55	31	.42	5.65	0.10
⟨pretty, beautiful⟩	Adjective (o)	8	11	73	50	.41	5.04	-0.46
⟨rare, extinct⟩	Adjective (c)	79	79	40	34	.29	5.83	1.05
⟨scarce, unavailable⟩	Adjective (c)	62	57	40	17	.18	4.78	0.29
⟨silly, ridiculous⟩	Adjective (o)	4	14	77	40	.43	4.17	0.01
⟨small, tiny⟩	Adjective (o)	4	25	80	27	.54	4.22	0.80
⟨snug, tight⟩	Adjective (o)	12	21	87	87	.30	2.86	-1.05
⟨some, all⟩	Quantifier	96	89	67	87	.79	5.83	-0.12
⟨sometimes, always⟩	Quantifier	100	86	80	90	.60	5.70	-1.05
⟨special, unique⟩	Adjective (c)	8	14	50	30	.32	3.48	0.54
⟨start, finish⟩	Verb	21	21	43	50	.40	4.95	0.70
⟨tired, exhausted⟩	Adjective (o)	4	14	57	41	.45	5.13	0.92
⟨try, succeed⟩	Verb	62	39	37	57	.35	5.82	1.23
⟨ugly, hideous⟩	Adjective (o)	4	18	37	31	.48	5.27	0.86
⟨unsettling, horrific⟩	Adjective (o)	29	25	37	37	NA	5.65	-0.48
⟨warm, hot⟩	Adjective (o)	75	64	70	38	.51	5.00	-0.28
⟨wary, scared⟩	Adjective (o)	21	14	40	37	.06	4.39	-0.48

Table 3: List of scales used in the experiments reported in this paper. Legend: **SI** *= percentages of participants that derived a scalar inference;* **Cloze** *= percentages of participants who mentioned a stronger scalar term in the modified cloze task (Exp. 3, lenient analysis);* +N *= neutral condition (Exp. 1);* –N *= non-neutral condition (Exp. 2);* **LSA** *= strength of association based on Latent Semantic Analysis (Footnote 6);* **Dist** *= mean difference in perceived semantic distance (Exp. 4);* **Freq** *= logarithm of the ratio between the frequency of the weaker scalar term and the frequency of the stronger scalar term (Section 4);* (c) *= closed scale;* (o) *= open scale.*

Comparison	β	**SE**	**Z**	***p***
Quantifier / Modal	0.75	0.46	1.65	.330
Quantifier / Verb	2.83	0.42	-6.77	.000
Quantifier / Adjective	2.89	0.39	7.42	.000
Modal / Adjective	2.13	0.27	8.01	.000
Modal / Verb	2.08	0.31	6.76	.000
Verb / Adjective	0.05	0.19	0.28	.992

Table 4: Comparisons between the percentages of positive answers per lexical category (Experiment 1).

The results of this pilot experiment correlated almost perfectly with the results from Experiment 1 ($r = .972, t(8) = 11.663, p < .001$). Apparently, the number of control items thus does not have a substantial effect on the outcome of the inference task.

Results

One participant was excluded from the analysis for making mistakes in three of the control items. Four out of a total of 1250 answers were missing. The results of the experiment are shown in Figure 2. It is evident from this graph that there was considerable variation among critical items, with positive responses ranging from 4% (for seven adjective scales) to 100% (for ⟨cheap, free⟩ and ⟨sometimes, always⟩).

All of the following analyses were carried out using R, a programming language and environment for statistical computing (R Development Core Team 2006). A coarse-level analysis suggests a pattern much like the one reported by Doran et al. (2009). We constructed a binomial mixed model using the `lme4` package (Bates & Maechler 2009), with the responses on the inference task as dependent variable, word class as independent variable, and participants and items as random factors (Barr, Levy, Scheepers, & Tily 2013). Multiple comparisons were carried out with Tukey's procedure using the `multcomp` package (Hothorn, Bretz, & Westfall 2008). The results of this analysis are given in Table 4. The proportion of positive answers for adjective scales (closed and open) ($M = .33$) was significantly lower than for modal expressions ($M = .80, p < .001$) or quantifiers ($M = .90, p < .001$), but it did not significantly differ from the proportion of positive answers for verbs ($M = .34, p = .992$). Verbs yielded significantly fewer positive answers than quantifiers ($p < .001$) and modal expressions ($p < .001$), while the difference between quantifiers and modal expressions was not significant ($p = .330$).

This global analysis might give the impression that there is a neat dichotomy

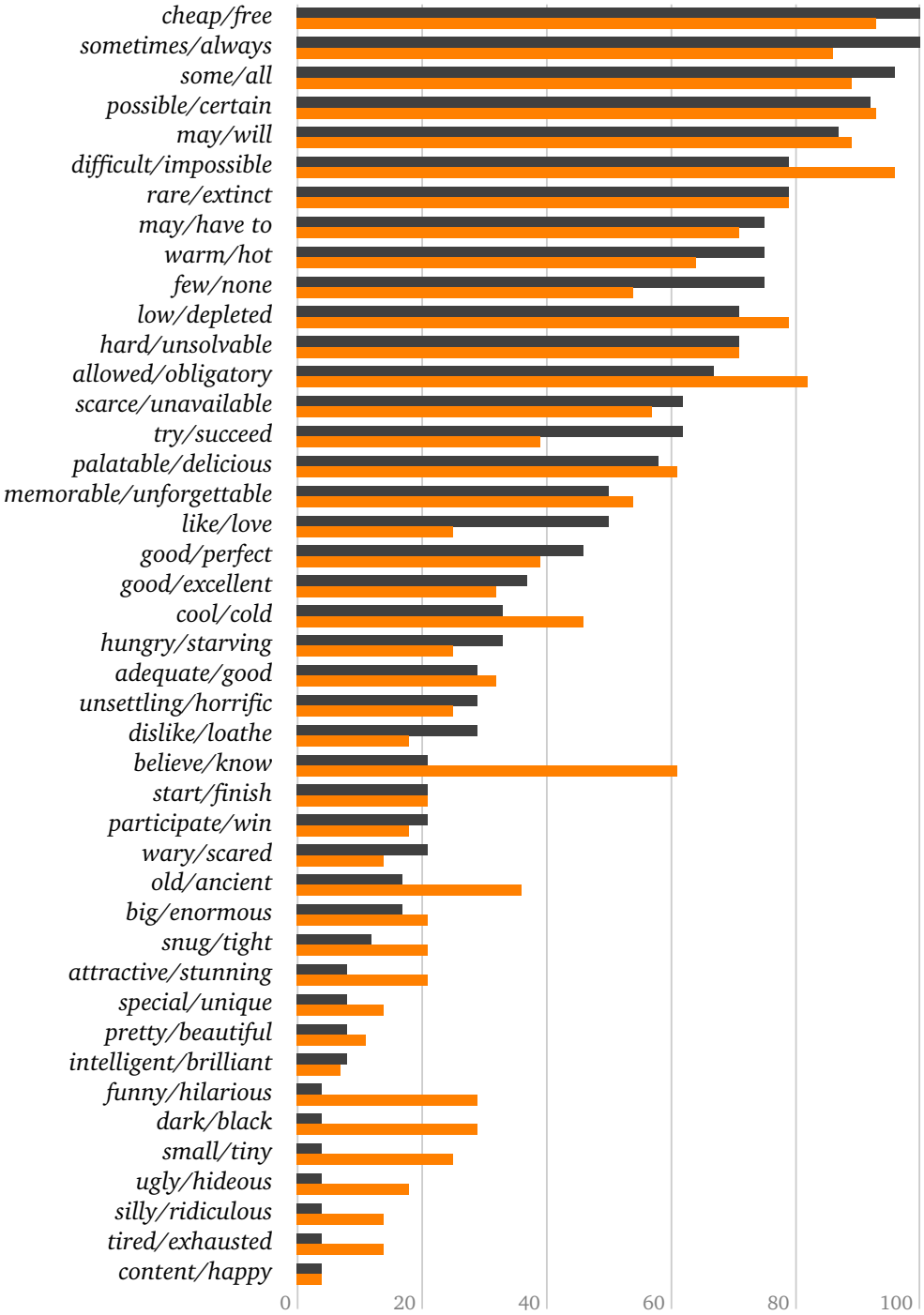

Figure 2: Percentages of positive responses in Experiment 1 (neutral content, dark grey) and Experiment 2 (non-neutral content, orange).

between adjectives and verbs, on the one hand, and quantifying and modal expressions on the other. However, a closer look at Figure 2 reveals that this conclusion is misleading. While quantifiers and modal expressions consistently gave rise to high proportions of scalar inferences, the results for adjectives and verbs were mixed. In the case of verbs, the difference between the lowest result (21% for ⟨believe, know⟩) and the highest result (62% for ⟨try, succeed⟩) was 41%; for the selection of adjectives, the difference was 96% with results varying from 4% (for seven scales) to 100% (for ⟨cheap, free⟩).

Generally speaking, while closed scales predominate in the upper half of the distribution shown in Figure 2 ($M = .62$), the lower half is populated mainly, though not exclusively, by open scales ($M = .23$). A binomial mixed model confirms that this difference is significant ($\beta = 2.327, SE = 0.472, Z = 4.936, p < .001$). It should be observed, however, that closed and open scales are not distributed evenly across the different grammatical categories. For instance, all of the quantifier and modal scales are closed, and we have seen that these categories were associated with relatively high proportions of positive responses. We therefore repeated the previous analysis, including only adjectival scales. Again, we found that closed scales ($M = .57$) yielded a significantly higher proportion of positive answers than open scales ($M = .21, \beta = 2.17, SE = 0.53, Z = 4.07, p < .001$). However, there wasn't a neat dichotomy between types of scale: on the one hand, some closed scales, notably ⟨dark, black⟩ ($M = .04$) and ⟨special, unique⟩ ($M = .08$) yielded a low proportion of positive answers; on the other hand, some open scales, notably ⟨warm, hot⟩ ($M = .75$), yielded a high proportion of positive answers.

Discussion

The results of our first experiment disprove the uniformity assumption: different scalar expressions yield markedly different rates of scalar inferences, ranging between 4% and 100%. Our results also disprove a weaker hypothesis, according to which scalar expressions of the same lexical category should pattern alike. Contrary to this hypothesis, we found that there is considerable variation in the behaviour of scalar expressions of the same category, especially adjectives. Hence, although broadly speaking, our findings agree with Doran et al.'s, they disagree in this point: whereas Doran et al.'s results suggest that the difference between quantifiers and adjectives is dichotomous, ours show that it is a matter of degree. For the reasons discussed in Section 1, we believe that our findings are more reliable.

There is one intriguing wrinkle in our data. In the foregoing, we mentioned that, in general, inference tasks give rise to higher levels of scalar inferences than verification tasks do. To some extent, we see this in the current study, too. In particular, existential quantifiers elicited scalar inferences more than 90% of the time, which is in line with Geurts and Pouscoulous' (2009) experiment 1, which

used the same task. This figure is substantially higher than in any verification study reported in the literature, including Doran et al.'s, who observed scalar inferences for quantifiers about 50% of the time. However, as a group, the adjectival items seem more or less impervious to this factor, with several of them scoring below 10%. Apparently, the effect of the inference task is not additive.

In this experiment, we used materials that were as neutral as possible, which was done mainly by using pronouns instead of complex noun phrases, but also by using generic predicates. One potential drawback of this approach is that it may have had a disorienting effect, leaving participants to wonder who or what these pronouns referred to, which, in its turn, may have affected our findings. To allay this worry, we replicated Experiment 1 with less neutral materials, using full noun phrases instead of pronouns, as well as more specific predicates. A further reason for doing this is that the results of this replication may clarify the extent to which rates of scalar inferences are influenced by sentence meaning and world knowledge. Although it is often taken for granted that the computation of scalar inferences is generally constrained by world knowledge (e.g., Franke 2011), Geurts (2010, 153) observes that, in some cases at least, the effect seems negligible, citing (7) as evidence:

(7) Cleo threw all her marbles in the swimming pool. Some of them sank to the bottom.

Even though it is highly implausible that some but not all of the marbles sank to the bottom, this is exactly what is implied by (7). Aside from serving as a control on the results of Experiment 1, then, Experiment 2 may also shed some light on this issue.

2.2. Experiment 2

Participants

We posted surveys for 30 participants on Amazon's Mechanical Turk (mean age: 32; range: 21–62; 14 females). Only workers with an IP address from the United States were eligible for participation. In addition, these workers were asked to indicate their native language. Payment was not contingent on their response to this question. Respondents were paid $0.70 for participating in this task, which took approximately 8 minutes. One participant was excluded from the analysis because she was not a native speaker of English. None of the participants in Experiment 2 had already participated in Experiment 1.

Materials

We tested the same scales as in Experiment 1, using the same procedure. However, in this case, the statements used by John and Mary contained more specific predicates and full noun phrases rather than pronouns. These statements were created on the basis of the following pretest. Ten participants (mean age: 35; range: 21–60; 6 females), all of them U.S. residents and native speakers of English, were drafted through Amazon's Mechanical Turk. Participants saw sentences containing a gap, like the following:

(8) a. The _____ is attractive but she isn't stunning.
    b. He is sometimes _____ but not always.

Statements always contained both the weaker and the stronger scalar term used in Experiment 1, so that any confusion about the meaning of the weaker scalar term was avoided. Otherwise, scalar terms like 'low' and 'hard', for instance, might have received an interpretation on which they are incompatible with 'depleted' and 'unsolvable', respectively. Participants were instructed to indicate how the blanks could be filled in so as to yield a natural-sounding sentence, and had to provide three completions for every statement.

Out of all the completions suggested by the participants in the pretest, we selected three per scale, applying two constraints. First, we sought to ensure sufficient variation for each scalar expression. To illustrate, in the case of (8a), we chose 'nurse', rather than 'singer', in addition to 'model' and 'actress'. Second, whenever possible, we selected two relatively frequent and one relatively infrequent completion for each scale; if the variation of suggested completions was too great to apply this criterion, a random selection was made. Thus we constructed three statements for every scale (see appendix). An example trial is given in Figure 3. Randomised lists were created for each participant, making sure that every statement was encountered by 10 participants (i.e. 1 in 3). Lastly, we included seven control items per list, in which the statement either entailed or was unrelated to the critical inference (see appendix).

Results

One participant was excluded from the analysis for making mistakes in four control items. Four out of a total of 1500 answers were missing. The results are given in Figure 2. Paired chi-square tests showed that only two scales yielded different rates of scalar inferences from Experiment 1, namely ⟨believe, know⟩, where the rate of positive responses increased from 20% to 60% ($\chi^2(1) = 7.42, p = .006$), and ⟨funny, hilarious⟩, where the rate of positive responses went from 4% to 30% ($\chi^2(1) = 4.05, p = .044$). Accordingly, the product-moment correlation between the

John says:

This student is intelligent.

Would you conclude from this that, according to John, she is not brilliant?

☐ Yes ☐ No

Figure 3: Sample item used in Experiment 2.

Condition	**Exp. 1**	**Exp. 2**	$\chi^2(1)$	p
Adjective	33	38	4.314	.038
Modal	81	85	0.479	.489
Quantifier	92	76	6.502	.011
Verb	34	30	0.542	.462

Table 5: Comparison between the proportions of positive responses in Experiments 1 and 2 for each of the four word classes.

proportions of positive answers for each item in the two experiments was quite high ($r = .909, t(41) = 13.98, p < .001$).

Table 5 compares the proportions of positive responses in Experiments 1 and 2 when scalar expressions of the same category are grouped. The mean percentage of positive responses for adjectives increased significantly from 33% in Experiment 1 to 38% in Experiment 2 ($p = .038$), while it decreased significantly from 92% to 76% in the case of the quantifiers ($p = .011$). The results for the other word classes did not differ significantly across Experiments 1 and 2. Overall, the rate of positive responses was 44%, which was not significantly higher than the 42% of positive responses in the first experiment ($\chi^2(1) = 0.85, p = .356$). Paired chi-square tests showed that there was no pair of statements for any scale that yielded significantly different rates of positive answers (though it should be noted that there were at most ten observations per statement).

The overall pattern of responses remained the same: the proportions of positive responses for quantifiers and modals was significantly higher than for verbs and adjectives, while none of the other comparisons yielded a significant result; see Table 6 for details.

Comparison	β	**SE**	**Z**	***p***
Quantifier / Modal	-0.70	0.36	-1.93	.206
Quantifier / Verb	2.08	0.31	-6.79	.000
Quantifier / Adjective	2.32	0.27	8.53	.000
Modal / Adjective	3.02	0.28	10.62	.000
Modal / Verb	2.78	0.32	8.77	.000
Verb / Adjective	0.24	0.20	1.21	.607

Table 6: Comparisons between the percentages of positive answers per lexical category (Experiment 2).

Discussion

Adding more content to the materials had a relatively small effect on the overall results, and did not affect the general conclusions we drew from the results of Experiment 1. This finding makes it highly unlikely that the outcome of that experiment can be attributed to lack of content, and it lends support to the view that, modulo the contribution of the scalar expression, sentence meaning and world knowledge have only a small effect on the probability that a scalar inference is derived.

Grouping together scalar expressions of the same category, we found that the proportion of positive responses for adjectives was significantly higher in Experiment 2 than in Experiment 1. This is in line with what Doran et al. (2009) found; they tested their items in three conditions, varying whether a competing scalar was mentioned in the question, and found that doing so only had an effect on adjectives, where it led to an elevated number of scalar inferences. In our experiments, too, adjectives yielded a higher rate of scalar inferences when they were presented in a richer context—although it should be noted that, just like in Doran et al.'s experiment, this effect was relatively modest (5% in our experiment). Unlike Doran et al.'s findings, we also found that the rate of scalar inferences for quantifiers decreased significantly when more sentential content was added. Apparently, these scalar expressions are sensitive to contextual cues as well.

Given our own data and Doran et al.'s, we can safely say that the uniformity assumption is false: the rates at which scalar expressions yield upper-bounding inferences fluctuate so wildly that it makes no sense to say that they are all more or less the same. How can we account this variability? We have already seen two factors that correlate with the pattern of responses observed in Experiments 1 and 2: open/closed syntactic class and open/closed scale. However, since these factors are dichotomous (and besides, in our materials, the latter subsumed the former), it is clear that they fall short of providing a full account of the data. Hence, in the

following sections, we discuss four further candidate factors, three of which we tested empirically.

3. Explaining diversity (1): Focus/non-focus

One possibility that is often mentioned in the literature is that the probability of a scalar inference is determined by the question under discussion (e.g., van Kuppevelt 1996, van Rooij & Schulz 2004, Zondervan 2010). On this view, a scalar expression will only give rise to an upper-bounding inference if it is (part of) the focus of an utterance. That is to say, B's answer in (9), but not in (10), should imply that she did not eat all of the cookies.

(9) A: How many cookies did you eat?
B: I ate [some]$_{\mathrm{F}}$ of the cookies.

(10) A: Who ate some of the cookies?
B: [I]$_{\mathrm{F}}$ ate some of the cookies.

Since in our experiments no questions were asked, a possible explanation for the differential ratings of sentences with (say) 'some' and 'small' is that participants tended to contextualise these sentences in different ways, with 'some' having a preference for the focus position, and 'small' having a preference for non-focus positions.

There are several reasons for being sceptical about this line of explanation. First, in our experiments, scalar adjectives always occurred in predicate position, which is widely agreed to be focused by default. Furthermore, in Experiment 1, grammatical subjects were always pronominal, and pronouns rarely receive focus; this is especially true of neuter 'it', but it holds for the other third-person pronouns, too. To illustrate, it is obvious that, in the following examples, the adjectives are most likely to be in focus:

(11) a. It is cheap.
b. It is small.

Still, whereas (11a) triggered scalar inferences in all cases, (11b) did so only 4% of the time.

Finally, the focus hypothesis has been extensively tested by Zondervan (2010), who conducted a series of experiments in which the (tacit or overt) question under discussion was manipulated in a variety of ways. Overall, the effects of these manipulations were either weak or nil. For example, in Zondervan's first experiment, (13) was presented as an answer to either (12a) (object focus) or (12b) (subject focus).

(12) a. What did Katja find?
b. Who found a crab or a starfish?

(13) Katja found a crab or a starfish.

In all conditions, the question/answer sequence was preceded by a vignette in which Katja actually found both a crab and a starfish, which falsified the scalar (exclusive) construal of the disjunction in (13). Participants had to indicate if they considered (13) a correct answer in this situation. Zondervan's finding was that, although the difference between (12a) and (12b) had an effect, the effect was quite small: whereas in the subject-focus condition, (13) was rejected 55% of the time, in the object-focus condition it was rejected 73% of the time. All the other experiments reported by Zondervan show the same picture: if the question under discussion had an effect on the derivation of scalar inferences at all, the effect was weak, at best.

Based on the foregoing considerations, we conclude that, although implicit focusing may have had some effect on the pattern of results obtained in Experiments 1 and 2, it is unlikely that this effect was very strong, and it is practically certain that, taken on its own, focusing cannot account for the extreme variation we observed. Therefore, we will go on to consider a number of alternative explanations.

4. Explaining diversity (2): Word frequencies

The first alternative explanation we will investigate empirically is focused on the word frequencies of the scalar expressions in the inference task. To see how these could play a role, we compare the scales ⟨some, all⟩ and ⟨big, enormous⟩, which gave rise to scalar inferences 96% and 16% of the time, respectively. It could be that this discrepancy was caused by the fact that, whereas 'all' is a quite common word that should be readily available to the speaker in a context in which he uttered 'some', 'enormous' is relatively rare, hence less likely to be readily available, which might explain why the speaker did not use it even if, strictly speaking, it was more appropriate than 'big'. In line with this proposal, a scale like ⟨warm, hot⟩, where the stronger scalar expression is more frequent than the weaker one, yielded scalar inferences in 72% of the cases. This explanation can be generalised and made more precise as follows:

Frequency Hypothesis:
Given a scale $\langle \alpha, \beta \rangle$, the probability that $\varphi[\alpha]$ licenses a scalar inference $\neg\varphi[\beta]$ is determined by the frequency of β relative to that of α.

In order to test this hypothesis, we extracted the frequencies of all scalar expressions in our materials from the Corpus of Contemporary American English (Davies

2008). For each scale, we divided the frequency of the stronger scalar term by the frequency of the weaker one, and logarithmised the outcome to reduce the skewness of the resulting distribution. The results of this analysis are given in Table 3. We constructed a binomial mixed model with the results of the inference task from Experiment 1 as dependent variable, relative frequencies as independent variable, and participants and items as random factors. There was no significant effect of the relative frequencies on the proportion of positive answers in the first inference task ($\beta = -0.061, SE = 0.345, Z = -0.177, p = .86$). In line with this result, the coarser product-moment correlation between the relative frequencies and the rates of positive answers was not significant, either ($r = .023, t(41) = .149, p = .88$).

An alternative possibility is that it is not relative frequency, but rather the bare frequency of the stronger alternative that determines the likelihood with which a scalar inference is derived. The idea would be that, even if 'horrific' is more frequent than 'unsettling', a speaker who uses 'unsettling' might not have considered 'horrific' simply because it is a rare word. To test this hypothesis, we carried out an analysis similar to the one reported in the last paragraph, but this time using logarithmised frequencies of the stronger scalar terms as predictor variable. Again, a mixed model with logarithmised frequencies of the stronger scalar terms as independent variable did not yield a significant effect ($\beta = 0.467, SE = 0.329, Z = 1.418, p = .156$), and the product-moment correlation was not significant, either ($r = .169, t(41) = 1.098, p = .278$).

The foregoing analyses were somewhat coarse-grained, because frequency distributions were not balanced across scales of different categories; nor would it have been feasible to do so. However, as explained in Section 2.1, we did balance the set of open adjectival scales by selecting ten scales in which the weaker scalar term was more frequent than the stronger one, and ten scales for which the opposite was true (Table 7). This subset of items confirmed the results of the previous analyses: once again, a binomial mixed model with the results of the inference task from Experiment 1 as dependent variable, relative frequencies as independent variable, and participants and items as random factors, showed no significant effect of relative frequency ($\beta = -0.446, SE = 0.358, Z = -1.248, p = .212$). An analysis with absolute instead of relative frequencies led to the same conclusion ($\beta = 0.753, SE = 0.483, Z = 1.186, p = .235$).

To sum up: it appears that neither the relative frequency of the scalar expressions nor the absolute frequency of the stronger term had a significant effect on whether or not a scalar inference is computed. We conclude, therefore, that frequency does not play a major role in the distribution of scalar inferences.

freq(α) > freq(β)	μ	**SI**	**freq(α) < freq(β)**	μ	**SI**
⟨silly, ridiculous⟩	0.01	4	⟨intelligent, brilliant⟩	-0.12	8
⟨attractive, stunning⟩	0.37	8	⟨cool, cold⟩	-0.21	33
⟨hungry, starving⟩	0.71	33	⟨warm, hot⟩	-0.28	75
⟨small, tiny⟩	0.80	4	⟨pretty, beautiful⟩	-0.46	8
⟨ugly, hideous⟩	0.86	4	⟨unsettling, horrific⟩	-0.48	29
⟨tired, exhausted⟩	0.92	4	⟨wary, scared⟩	-0.48	21
⟨old, ancient⟩	1.08	17	⟨content, happy⟩	-0.85	4
⟨big, enormous⟩	1.13	17	⟨palatable, delicious⟩	-0.89	58
⟨funny, hilarious⟩	1.17	4	⟨snug, tight⟩	-1.05	12
⟨good, excellent⟩	1.34	37	⟨adequate, good⟩	-1.52	29

Table 7: Logarithms of the relative frequencies (μ) for two sets of ten open adjective scales $\langle\alpha,\beta\rangle$, where α was either more (left column) or less frequent (right column) than β. **SI** *= rates of scalar inferences in Experiment 1.*

5. Explaining diversity (3): Association

It is often said that scalar terms 'evoke' a scale: when you hear a sentence containing 'some', its scalemate 'all' immediately comes to mind, which explains why a scalar inference is so readily computed. Perhaps, then, the reason why 'small' very rarely implies 'not tiny' is that 'small' normally does not evoke the scale ⟨small, tiny⟩. In other words, someone who hears 'small', as opposed to 'some', does not automatically consider the stronger scalar expression, and is therefore less likely to ask herself why the speaker did not use it instead, which explains the near-absence of a scalar inference. Thus we arrive at the following hypothesis:

Association Hypothesis:
Given a scale $\langle\alpha,\beta\rangle$, the probability that $\varphi[\alpha]$ licenses a scalar inference $\neg\varphi[\beta]$ is determined by the strength of assocation between α and β.

In order to test this hypothesis, we need a measure of the strength of association between two scalar expressions. To this end, we conducted a modified cloze task. A standard cloze test consists of sentences or text fragments with certain words removed, where participants are asked to replace the missing words. We modified this design by underlining instead of removing words. Participants were asked to list three alternatives to a given sentence $\varphi[\alpha]$ by replacing the underlined scalar term α with whatever expression they saw fit. We assumed that the stronger the association between α and β, the more likely it would be that participants replaced α with β.

She is intelligent.

She is _____
She is _____
She is _____

Figure 4: Sample item used in Experiment 3.

5.1. Experiment 3

Participants

We posted surveys for 60 participants on Amazon's Mechanical Turk (mean age: 36; range: 21–57; 21 females). Only workers with an IP address from the United States were eligible for participation. In addition, these workers were asked to indicate their native language. Payment was not contingent on their response to this question. Respondents were paid $1.25 for participating in this task, which took approximately 15 minutes. All participants were native speakers of English. Two of the participants had already participated in Experiment 1 or 2. We included these participants in the analysis we discuss below. Excluding them would not change the statistical significance of any of the p-values we report.

Materials and procedure

Figure 4 shows an example of a critical item. Each trial consisted of a sentence with a scalar term that was underlined. Participants were instructed to indicate which words could have occurred instead of the underlined word. Half of the participants saw the neutral statements used in Experiment 1; the other half saw the non-neutral statements from Experiment 2. We constructed two minimally different sets of instructions. One version is given below:[5]

> In the following you will see 43 sentences. In every sentence, one word will be highlighted, like this:
>
> She is angry.

5. Note that the neutral version included only 41 statements, the reason being that the statements for ⟨good, excellent⟩ and ⟨good, perfect⟩, on the one hand, and ⟨may, have to⟩ and ⟨may, will⟩, on the other, were identical in this version of the task. In the analysis reported on below, we paired the results for these statements with the results on the inference task for ⟨good, excellent⟩ and ⟨may, have to⟩, respectively. However, changing this pairing did not have any effect on the results.

> Which words could have occurred instead of the highlighted one? Some of the alternatives that may come to mind are *beautiful*, *happy*, *married*, and so on. We ask you to tell us the first three alternative words that occur to you when you read these sentences. We are interested in your spontaneous responses, so don't think too long about it.

In the second version, the first sample alternative (here 'beautiful') was replaced with a scalar term that was stronger than the highlighted expression (namely 'furious'). We did this to control for the possibility that mentioning or not mentioning a stronger expression in the instructions might have an effect on the participants' responses. More precisely, participants might be more likely to provide stronger scalemates if a stronger scalemate had been mentioned in the instructions. A different list was constructed for each of the participants, varying the order of the trials.

Results and discussion

Seven out of a total of 2550 answers were missing. We annotated our results in two different ways. For each trial, we first coded if the participant mentioned the stronger scalar term we used in the inference tasks. However, this measure might be too strict because, first, there were scalar terms for which participants often mentioned a synonymous or nearly synonymous stronger scalar term, like 'every' instead of 'all'. Secondly, it might be the case that participants compute a scalar inference if they consider any stronger scalar term, even if it is different from the one in the question. For instance, a participant who associates 'possible' with 'probable', and computes a scalar inference on the basis of the scale ⟨possible, probable⟩, thereby also infers that it is not certain, even though she did not consider that particular alternative. Therefore we also determined for each trial whether any stronger scalar term was mentioned.

	–N	+N
+S	25	29
–S	18	26

Strict coding

	–N	+N
+S	48	52
–S	41	48

Lenient coding

Table 8: Percentages of responses in Experiment 3 which mentioned either the same scalar term we used in our inference tasks (Strict coding) *or any stronger scalar term* (Lenient coding). *Instructions either contained a stronger scalar term* (+S) *or not* (–S)*, and sentences were neutral* (N) *or not* (–N).

The results of our analyses are summarised in Table 3. We start with the strict coding scheme. We first conducted a loglinear analysis to test whether the probability

that the stronger scalar term used in the inference task was mentioned was affected by (a) whether or not the target sentences were neutral (+N vs. –N) and (b) whether or not a stronger scalar expression was mentioned in the instructions (+S vs. –S). A summary of the effects of these factors is given in Table 8. Overall, the stronger scalar term was mentioned in 25% of the trials. It was mentioned significantly more often with neutral statements (27%) than with non-neutral ones (22%, $G^2(1) = 11.61, p < .001$). However, this effect interacted with the form of the instructions ($G^2(2) = 13.74, p = .001$): it was only significant if the instructions did not contain a stronger scalar term ($G^2(1) = 12.28, p < .001$). The stronger scalar term was also mentioned significantly more often when the instructions contained a stronger scalar term (27%) than when they did not (22%, $G^2(1) = 7.28, p < .01$), and again there was an interaction with the neutral/non-neutral factor ($G^2(2) = 10.46, p = .005$): the effect reached significance for non-neutral statements only ($G^2(1) = 9.12, p < .005$).

Presumably, the reason why stronger scalar terms were mentioned more often in the neutral condition is that in this condition, the scalar term was more or less the only thing to go on, whereas in non-neutral condition, associations were constrained by the sentential context as well. To illustrate, compare the following sentences:

(14) a. That house is old.
b. It is old.

Whereas in the case of (14a) participants were likely to mention properties they associated with houses or old houses, (14b) is obviously much less constraining. Mentioning a stronger scalar term in the instructions dampened this effect.

With the lenient coding scheme, we found a very similar pattern. A stronger scalar term was mentioned in 47% of the trials. It was mentioned significantly more often with neutral than non-neutral sentences (50% vs. 45%, $G^2(1) = 5.85, p < .025$). As with the strict coding scheme, this effect interacted with the form of the instructions ($G^2(2) = 6.64, p < .05$): it only reached significance if the instructions did not contain a stronger scalar term ($G^2(1) = 5.26, p < .025$). Stronger scalar terms were mentioned significantly more often if the instructions contained a stronger scalar term than when they did not (50% vs. 44%, $G^2(1) = 7.28, p < .01$). There was an interaction with the neutral/non-neutral factor: the effect was only significant with non-neutral statements ($G^2(1) = 6.64, p < .01$).

Let us now examine the Association Hypothesis in light of the foregoing results. We constructed a binomial mixed model with the responses in the corresponding inference task as dependent variable. The independent variable was the proportion of times participants in Experiment 3 mentioned either the same stronger scalemate that was presented in the inference tasks ('SAME-response', strict coding) or any stronger scalar term ('ANY-response', lenient coding). Participants and items were

included as random factors. For neutral sentences, there was a marginally significant negative effect of the proportion of SAME-responses on the likelihood of a positive answer in the inference task ($\beta = -2.438, SE = 1.311, Z = -1.859, p = .063$), and no significant effect of ANY-responses ($\beta = -1.707, SE = 1.311, Z = -1.302, p = .193$). No significant effects were observed for non-neutral sentences (strict coding: $\beta = -0.975, SE = 1.327, Z = -0.735, p = 0.462$; lenient coding: $\beta = 0.374, SE = 1.187, Z = 0.315, p = 0.753$).

Since utterances are normally interpreted in a context, it seems reasonable to suppose that the results of the modified cloze task using non-neutral sentences are a more reliable indicator of how likely a particular (type of) expression is to come to mind if a scalar expression is encountered. Be that as it may, our results fail to support the Association Hypothesis. Whether or not a scalar inference is computed, for a given sentence $\varphi[\alpha]$, where α is a scalar expression, does not seem to depend on the strength of the association between α and any stronger alternative. For example, in the case of 'snug', nearly all participants in Experiment 3 mentioned 'tight' as an alternative, but in Experiments 1 and 2 the average rate of the snug/not-tight inference was a mere 16.5%; similar observations hold for ⟨pretty, beautiful⟩ and ⟨dislike, loathe⟩. On the other hand, there was a substantial group of scales that yielded high rates of scalar inferences, but for which stronger scalar terms were rarely mentioned in Experiment 3, clear examples being ⟨cheap, free⟩, ⟨hard, unsolvable⟩ and ⟨difficult, impossible⟩. In sum, the findings of this experiment argue against the hypothesis that rates of scalar inferences are determined by the strength of the connections between weaker and stronger scalar terms.[6]

6. *Explaining diversity (4): Distance*

The final explanation we will consider in this paper was inspired by an observation by Horn (1972, 90). Consider the following examples:

6. Chris Cummins (p.c.) noted that association strength can also be measured using a corpus-based method by means of latent semantic analysis (Landauer & Dumais 1997). LSA constructs a matrix with sentences from the corpus as columns and words that occur in these sentences as rows. A row consists of binary values that represent whether the word in question occurs in the sentence denoted by the column; so words that co-occur in a sentence have a 1 in the same column. Words that are closely related are expected to occur relatively often with the same words. Based on this matrix, LSA computes a measure of the association between different words.
On the basis of Landauer, Foltz, and Laham's (1998) LSA implementation, we obtained similarity values for each pair of scalar terms through pairwise, term-to-term comparisons with "general reading up to first year of college" as topic space. These similarity values, which are provided in Table 3, were used as an estimator of the results on the first inference task in a binomial mixed model with participants and items as random factors. However, we did not find a significant effect ($\beta = 2.565, SE = 1.660, Z = 1.545, p = .122$). The product-moment correlation was not significant either ($r = .233, t(40) = 1.514, p = .138$).

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

It seems to us that, normally speaking, an utterance of (15a) would be more likely to implicate the negation of (15c) than the negation of (15b). The intuitive reason for this divergence is that the difference in semantic strength between (15a) and (15c) is greater than that between (15a) and (15b). This intuition was partially vindicated by Zevakhina (2012), who found that participants considered the semantic distance between 'some' and 'all' greater than between 'some' and 'most'; although she did not find such a pattern for adjectival scales like ⟨warm, hot, sweltering⟩. The idea underlying the following hypothesis is that the highly variable rates at which scalar inferences are drawn might be explained in terms of the semantic distance between the weaker and the stronger term:

Distance Hypothesis:
Given a scale $\langle \alpha, \beta \rangle$, the probability that $\varphi[\alpha]$ licenses a scalar inference $\neg\varphi[\beta]$ is determined by the semantic distance between α and β.

Obviously, this hypothesis presupposes that it makes sense to compare pairs of expressions from different scales, and thus requires an absolute measure of semantic distance. Assuming that there is such a thing and that speakers have reliable intuitions about it (and neither assumption seems entirely unreasonable to us), the Distance Hypothesis leads us to expect that speakers' intuitions about semantic distance should at least be a partial predictor of the pattern of responses we observed in Experiments 1 and 2. Therefore, we conducted an experiment in which participants were asked, for all scales $\langle \alpha, \beta \rangle$ used in Experiments 1 and 2, how much stronger $\varphi[\beta]$ is relative to $\varphi[\alpha]$, and compared the results to the findings of those experiments.

6.1. Experiment 4

Participants

We posted surveys for 25 participants on Amazon's Mechanical Turk (mean age: 33; range: 20–62; 15 females). Only workers with an IP address from the United States were eligible for participation. In addition, these workers were asked to indicate their native language. Payment was not contingent on their response to this question. Respondents were paid $0.80 for their participation, which took approximately 15 minutes. One participant was excluded from the analysis because she was not a native speaker of English. Two participants had also participated in Experiment 1 or 2. We included these participants in the analysis. Excluding them would not change the statistical significance of any of the p-values we report.

Materials and procedure

An example trial is given in Figure 5. Participants were instructed to indicate whether and, if so, to what extent a statement with the higher-ranked scalar term was stronger than the same statement with the lower-ranked scalar term, by selecting a value on a seven-point scale. The instructions went as follows:

> Consider the following claims:
>
> 1. This is okay.
> 2. This is fantastic.
>
> Clearly, claim 2 is stronger than claim 1.
> Now compare the following claims:
>
> 3. This is fantastic.
> 4. This is marvelous.
>
> Here, neither claim seems much stronger than the other, if they differ in strength at all. In this questionnaire, we will show you a number of sentence pairs like the ones above. In each case, we ask you to indicate on a 7-point scale how much stronger the second claim is, where 1 means that the two claims are equally strong, and 7 means that the second claim is much stronger than the first one.

For this test, the neutral statements of Experiment 1 were used. Different lists of items were constructed for all participants, varying the order of the trials. Seven control items were included, each of which involved two statements which were synonymous or nearly so. These control statements involved the following pairs of words: *enormous/immense*, *fantastic/sensational*, *gifted/talented*, *obvious/clear*, *unbearable/intolerable*, *unexpected/unforeseen*, and *unpleasant/disagreeable*.

---

1. She is intelligent.
2. She is brilliant.

Is statement 2 stronger than statement 1?

equally strong	○ 1	○ 2	○ 3	○ 4	○ 5	○ 6	○ 7	*much stronger*

Figure 5: Sample item used in Experiment 4.

Results and discussion

Eight out of a total of 1250 answers were missing. One participant was excluded from the analysis because her mean rating for the control items exceeded two standard deviations from the grand mean for these items. The results of the experiment are presented in Table 3. The average rating for the target items was 5.12, as opposed to 2.81 for the synonymous control items. We first investigated whether the distance between scalar expressions was affected by their grammatical category or by their occurrence on a closed or open scale, by creating mixed models with these factors as independent variables. There were no significant differences in semantic distance between grammatical categories, except for the difference between adjectival ($M = 4.90$) and verbal ($M = 5.63$) scales ($\beta = 1.335, SE = 0.387, t = 3.450, p < .001$). The difference between closed ($M = 5.307$) and open ($M = 4.913$) scales was marginally significant ($\beta = 0.394, SE = 0.217, Z = 1.818, p = .069$). These results suggest that the semantic distance between scalar expressions is largely independent of their grammatical category or their occurrence on a closed or open scale.

In order to establish whether the results on the distance task were a good approximation of the results on the inference task, we constructed a binomial mixed model with the responses on the inference task in Experiment 1 as dependent variable, the mean ratings on the distance task as independent variable, and participants and items as random factors. There was a significant effect of the results of the distance task on the results of the inference task ($\beta = 1.335, SE = 0.387, t = 3.450, p < .001$).

This finding confirms the prediction made by the Distance Hypothesis, though it must be noted that semantic distance only accounts for a relatively small part of the inference data, since a simple product-moment correlation test shows that the variance explained by the Distance Hypothesis is 20% ($r = .449, t(41) = 3.223, p = .002$). This leaves a rather large amount of variation unaccounted for.

At this point, we constructed a mixed model with all of the factors we discussed so far (word class, frequency, strength of association, and semantic distance) as predictors of the rates of scalar inferences we found in Experiment 1. This allows for the possibility that one of these factors only affects the rates of scalar inferences when the effect of another factor is accounted for. The class of open-ended adjectives was included as reference category. The results of this mixed model, summarised in Table 9, confirmed our previous analyses: only word class and semantic distance had a significant effect on the rates of scalar inferences

Parameter	β	**SE**	**Z**	***p***
Distance	0.84	0.13	6.49	.000
Frequency	-0.11	0.09	-1.28	.202
Association	-0.16	0.35	-0.47	.640
Adjective (c)	1.70	0.22	7.84	.000
Modal	2.63	0.30	8.72	.000
Quantifier	3.38	0.46	7.33	.000
Verb	0.22	0.23	0.98	.326

Table 9: Parameters of a mixed model with the results from Experiment 1 as dependent variable, the logarithms of the ratio between the frequency of the weaker scalar term and the frequency of the stronger scalar term (Section 5), the mean differences in perceived semantic distance (Exp. 4), the percentages of participants who mentioned a stronger scalar term in the modified cloze task (Exp. 3, lenient analysis), and word class as independent variables, and participants as random variable.

7. General discussion and conclusion

Until recently, neither the experimental nor the theoretical literature on scalar inferences showed much concern for the diversity of scalar expressions, and by and large confined their attention to less than a handful of items, notably 'some' and 'or'. Apparently, the tacit assumption was that these are representative of a much larger family of scalar terms. That assumption turns out to be mistaken: following up on studies by Doran et al. (2009) and Larson et al. (2009), we have shown that the rates at which scalar expressions give rise to upper-bounding inferences could hardly be more diverse, and that the ⟨some, all⟩ scale, which has been the workhorse of recent research on scalar inferences, is an extreme case (Experiments 1 and 2).

While this is our main finding, a large part of the foregoing discussion addressed the question of how the observed diversity can be accounted for. In a series of follow-up experiments, we looked for factors that might help to explain the results of Experiments 1 and 2. We found evidence for three such factors, and failed to find evidence for another three. The factors for which we did find supporting evidence are the following:

i. Open vs. closed class (Experiments 1 and 2)
Overall, closed-class scalars (quantifiers and modal verbs) gave rise to upper-bounded inferences more often than open-class scalars (adjectives and verbs). All closed-class items ranged in the upper third of the ranking shown in Figure 2, though it bears emphasising that this region contained many open-class items,

too, and that, strictly speaking, the comparison between closed-class and open-class items is somewhat dubious, since the latter greatly outnumbered the former. Furthermore, since all the closed-class items involved scalar expressions that denote endpoints on closed scales, this factor is not independent from the following one.

ii. *Open vs. closed scales* (Experiments 1 and 2)
Scalar expressions that inhabit a closed scale (i.e., a scale with an end point) were more likely to give rise to upper-bounded inferences than their open-scale counterparts. While closed scales predominate in the upper half of the distribution in Figure 2, the lower half is populated mainly by open scales. However, there is no strict dichotomy: inference rates were high for some open scales, too, and low for some of the closed scales.

iii. *Semantic distance* (Experiment 4)
The difference in strength between $\varphi[\alpha]$ (e.g. 'It is warm') and $\varphi[\beta]$ (e.g. 'It is hot') showed a positive correlation with the likelihood that $\varphi[\alpha]$ would trigger the inference that $\neg\varphi[\beta]$. However, this correlation wasn't very strong, and left 80% of the variation observed in Experiments 1 and 2 unaccounted for.

The three factors that we tested but failed to obtain evidence for are the following:

i. *Specificity* (Experiments 1 and 2)
The only difference between the target sentences used in Experiments 1 and 2 was that in the latter study sentences were more specific than in the former (e.g., 'It is warm' vs. 'The sand is warm'). This manipulation had essentially no effect: the correlation between the rates of upper-bounding inferences in the two experiments was .91.

ii. *Word frequencies* (Section 4)
The probability that $\varphi[\alpha]$ gives rise to the inference that $\neg\varphi[\beta]$, where β is a stronger scalemate of α's, might be correlated with the frequency of β. We tested two versions of this idea, measuring β's frequency either in absolute terms or relative to α's frequency, but neither version was supported by the data.

iii. *Association strength* (Experiment 3)
The probability that $\varphi[\alpha]$ gives rise to the inference that $\neg\varphi[\beta]$, might be correlated with the association strength between α and β (relative to the sentence frame $\varphi[\ ]$) or with the association strength between α and any other stronger scalemate of α's. However, we didn't find evidence for this hypothesis, either.

What are we to do about these findings? As far as we can tell, there is currently no theory on the market that can account for the response pattern we observed in Experiments 1 and 2. True, most theories are consistent with our findings,

but that's because they fail to make predictions about the variability of scalar inferences.[7] Therefore, instead of discussing individual theories, we will consider a number of ideas that underlie many or most extant theories of scalar inference, and consider whether any of them might contribute to a better understanding of the data presented in the foregoing.

It is widely held that the derivation of scalar inferences is contingent upon the availability of alternatives: stronger sentences the speaker could have uttered instead of the one he did utter.[8] Where do these alternatives come from? The standard answer, due to Horn (1972), is that alternatives are generated on the basis of lexical scales. Not to put too fine a point on it, the idea is that, if a speaker utters a sentence $\varphi[\alpha]$, and there is a scale $\langle \alpha, \beta \rangle$, then $\varphi[\beta]$ is an alternative. This approach has its limitations (Sauerland 2004, Katzir 2007, Geurts 2010), the most pertinent of which, for our current purposes, is that it is unenlightening, because it requires that we deal with scalar inferences on a case-by-case basis. Yet, the data presented in the foregoing might be construed as suggesting that that is precisely what is needed to account for scalar inferences; we will come back to this point.

Katzir (2007) presents a method for generating alternatives that does away with scales and has the additional merits of being simpler and more general than the scale-based approach (see also Fox & Katzir 2011). The core of Katzir's proposal is to derive alternatives from the speaker's sentence by deleting parts of it and/or exchanging words. For (16), this procedure yields, inter alia, the alternatives in (17):

(16) Wilma fed the ducks or some of the geese.

(17) a. Wilma fed the ducks.
b. Wilma fed some of the geese.
c. Wilma fed none of the geese.
d. Wilma fed all the geese.
e. Wilma fed the ducks and some of the geese.

Upper-bounding inferences are then derivable from those alternatives in (17) that are more informative than (16). Attractive though this method may be, it obviously doesn't help to account for the variability of scalar inferences, for the simple reason that, as far as this component of Katzir's analysis is concerned, there is no principled way of discriminating between alternatives so that some scalar inferences come out

7. An exception has to be made for theories which hold that scalar inferences are defaults (e.g., Gazdar 1979, Levinson 2000), but such theories seem committed to the prediction that all scalar inferences are reasonably frequent, which is inconsistent with our data, as well as many other data that have been reported in the literature. See Geurts (2010, Chapter 5) for discussion.

8. On some accounts, alternatives need not be stronger than what the speaker said: it suffices if they are at least as strong (e.g., Spector 2003).

being more likely than others.

It appears that, taken on its own, neither method for generating alternatives brings us closer to a better understanding of the variability of scalar inferences. This is not so surprising, since both methods use the same ingredients for generating the alternatives that produce scalar inferences: the speaker's sentence, a set of lexical substituents, and logical strength. It is natural to suppose that an explanatory account of scalar diversity will have to say more about the content of scalar expressions and their carrier sentences. The extant literature suggests two directions that such an explanation might take.

Starting with Soames (1982), scalar inference has often been treated as a two-step process, along the following lines (e.g., Sauerland 2004, van Rooij & Schulz 2004, Geurts 2010). Let ψ be a stronger alternative to φ. If speaker S utters φ, the first inference step is that it is not the case that S believes that ψ is true: $\neg \mathrm{Bel}_S \psi$. This is weaker than what is usually called a scalar inference, which is of the form $\mathrm{Bel}_S \neg \psi$. However, the stronger inference follows from the weaker one if S is 'competent' (or 'knowledgeable' or 'opinionated') with respect to ψ, which is to say that $\mathrm{Bel}_S \psi \vee \mathrm{Bel}_S \neg \psi$.

The two-stage model of scalar inference suggests the possibility that the differential rates of scalar inferences are due to the fact that the plausibility of the competence assumption varies from case to case. If this is correct, the reason why 'It is cheap' produced significantly more scalar inferences than 'It is small' would be that our participants considered it much more likely that the speaker was competent with regards to the proposition that it is free than with regards to the proposition that it is tiny. We don't find this line of explanation particularly promising, though. Take the sentence 'She is pretty', for instance. It seems to us that a speaker who utters this sentence will typically have an opinion as to whether the person in question is beautiful or not, and yet the sentence prompted a positive response only 8% of the time. Since this is not an isolated example, we are inclined to believe that competence is not the key.

Thus we are left with one last resort: relevance. Relevance is generally presupposed in theories of scalar inference (it is also built into Katzir's, for example), though it is rarely called upon to do explanatory work. In the scalar literature, relevance is usually understood in a narrow sense, as relevance to the question under discussion, and as we argued in Section 3, there are various reasons to doubt that relevance, thus understood, was a key factor in Experiments 1 and 2. However, although we don't wish to contest that the question of discussion captures an important aspect of relevance, there are other aspects as well. Geurts (2010, 44-50) distinguishes between three main sources of relevance: the relevance of a given piece of information may be measured in terms of (a) its contribution to the discourse ('question under discussion' is a special case of this), (b) the hearer's interests, or (c) general interest. Here we want to focus on the last category.

To explain the notion of general interest, consider the following minimal pair (Geurts 2010, 48):

(18) a. I saw an animal on the lawn this morning.
b. I saw a dog on the lawn this morning.

Whereas (18a) is prone to imply that the speaker is not sure what species of animal he saw, it would require special circumstances for (18b) to convey that the speaker is not sure what breed of dog he saw. Relatedly, whereas in English, a sentence like (19) will rarely convey that the speaker does not know whether Michio is older or younger than Takashi, its literal translation into Japanese standardly carries this suggestion (Matsumoto 1995):

(19) Michio is Takashi's brother.

The contrast between English and Japanese is partly explained by the fact that, in addition to words corresponding to English 'brother' and 'sister', Japanese also has sibling terms that encode gender as well as seniority. However, this is only part of the story, for we also have to suppose that, in Japanese, the hyponyms of 'brother' and 'sister' are the preferred means for introducing siblings.

The upshot of these observations is that 'there are general expectations to the effect that, when introducing a new discourse entity, speakers should employ expressions of at least a minimum level of specificity: 'sofa' and 'dog' are sufficiently specific; 'piece of furniture' and 'animal' are not. These expectations are general in that they don't fluctuate wildly from context to context or from hearer to hearer, though they may be culture specific.' (Geurts 2010, 48)

We believe that the data of our Experiments 1 and 2 are naturally viewed in terms of general interest. Whereas the difference between 'cheap' and 'free', 'sometimes' and 'always', or 'rare' and 'extinct' will be of interest in pretty much any context in which these expressions are used, the same does not hold for 'pretty'/'beautiful', 'silly'/'ridiculous', or 'tired'/'exhausted'. Needless to say, these judgements are not entirely impervious to contextual influences, but they appear to be relatively stable.

If this explanation is on the right track, many of the findings reported in the foregoing fall into place. Most importantly, if the response patterns observed in Experiments 1 and 2 are shaped by considerations of general interest, it is only to be expected that scalar inferences should are endorsed at varying rates; which is what we found. Furthermore, if an expression denotes the endpoint of a closed scale, it will be a salient alternative to its weaker scalemates, and therefore it makes sense that such expressions should trigger scalar inferences more often than expressions that do not denote endpoints. Similarly for semantic distance: the greater the difference in strength between α and β, the more likely it is that $\varphi[\beta]$ will be considered a relevant alternative to $\varphi[\alpha]$.

Admittedly, this explanation is largely programmatic, and a lot of work will have to be done to test its viability. In particular, the key concept of 'general interest' requires more attention and care than it has received so far. However, we believe that the general idea is a promising one.

The main purpose of this paper was to present data showing that scalar inferences exhibit a variability that poses a challenge for all extant theories, be they pragmatic, lexical, or syntactic. The same holds for the experimental research that has burgeoned since Bott and Noveck's (2004) landmark paper. If scalar diversity had been taken more seriously, especially in the various debates of the last decade, the current landscape of issues, theories, and experimental data would have looked quite different from what we are seeing now.

References

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition, 118*, 84–93.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bates, D., & Maechler, M. (2009). lme4: Linear mixed-effects models using S4 classes. R package version 0.999375-32. URL: `http://CRAN.R-project.org/package=lme4`.

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language, 66*, 123–142.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: the onset and time course of scalar inferences. *Journal of Memory and Language, 51*, 437–457.

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalized scalar implicatures generated by default? An online investigation into the role of context in generating pragmatic inferences. *Cognition, 100*, 434–463.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5.

Chevallier, C., Noveck, I. A., Nazir, T., Bott, L., Lanzetti, V., & Sperber, D. (2008). Making disjunctions exclusive. *The Quarterly Journal of Experimental Psychology, 61*(11), 1741–1760.

Davies, M. (2008). The Corpus of Contemporary American English: 450 million words, 1990 – present. URL: `http://corpus.byu.edu/coca/`.

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: dual task impact on scalar implicature. *Experimental Psychology*, *54*, 128–133.

Doran, R., Baker, R., McNabb, Y., Larson, M., & Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *International Review of Pragmatics*, *1*, 1–38.

Feeney, A., Scrafton, S., Duckworth, A., & Handley, S. (2004). The story of *some*: everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology*, *58*(2), 121–132.

Fox, D., & Katzir, R. (2011). On the characterization of alternatives. *Natural Language Semantics*, *19*, 87–107.

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics & Pragmatics*, *4*(1), 1–82.

Gazdar, G. (1979). *Pragmatics: implicature, presupposition, and logical form*. New York: Academic Press.

Geurts, B. (2010). *Quantity implicatures*. Cambridge: Cambridge University Press.

Geurts, B., & Pouscoulous, N. (2009). Embedded implicatures?!? *Semantics & Pragmatics*, *2*(4), 1–34.

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: evidence for immediate pragmatic enrichment. *Cognition*, *116*, 42–55.

Guasti, M. T., Chierchia, G., Crain, S., Foppolo, F., Gualmini, A., & Meroni, L. (2005). Why children and adults sometimes (but not always) compute implicatures. *Language and Cognitive Processes*, *20*(5), 667–696.

Hirschberg, J. (1991). *A theory of scalar implicature*. New York: Garland Press.

Horn, L. R. (1972). *On the semantic properties of logical operators in English*. Ph.D. thesis, UCLA. Distributed by Indiana University Linguistics Club.

Horn, L. R. (2007). Toward a Fregean pragmatics: *Voraussetzung*, *Nebengedanke*, *Andeutung*. In I. Kecskes, & L. R. Horn (Eds.) *Explorations in pragmatics: linguistic, cognitive and intercultural aspects*, (pp. 39–69). New York, NY: Mouton de Gruyter.

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal*, *50*(3), 346–363.

Huang, Y. T., & Snedeker, J. (2009). On-line interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*, 376–415.

Katzir, R. (2007). Structurally-defined alternatives. *Linguistics and Philosophy*, *30*(6), 669–690.

van Kuppevelt, J. (1996). Inferring from topics: scalar implicatures as topic-dependent inferences. *Linguistics and Philosophy*, *19*, 393–443.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259–284. URL: `http://lsa.colorado.edu/`.

Larson, M., Doran, R., McNabb, Y., Baker, R., Berends, M., Djalili, A., & Ward, G. (2009). Distinguishing the said from the implicated using a novel experimental paradigm. In U. Sauerland, & K. Yatsushiro (Eds.) *Semantics and pragmatics: from experiment to theory*, (pp. 74–93). Berlin: Palgrave MacMillan.

Levinson, S. C. (2000). *Presumptive meanings: the theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Matsumoto, Y. (1995). The conversational condition on Horn scales. *Linguistics and Philosophy*, *18*(1), 21–60.

Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, *78*, 165–188.

Noveck, I. A., Chierchia, G., Chevaux, F., Guelminger, R., & Sylvestre, E. (2002). Linguistic-pragmatic factors in interpreting disjunction. *Thinking and Reasoning*, *8*(4), 297–326.

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: an evoked potentials study. *Brain and Language*, *85*, 203–210.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition*, *78*, 253–282.

Pijnacker, J., Hagoort, P., van Buitelaar, J., Teunisse, J.-P., & Geurts, B. (2009). Pragmatic inferences in high-functioning adults with autism and Asperger syndrome. *Journal of Autism and Developmental Disorders*, *39*, 607–618.

Pouscoulous, N., Noveck, I. A., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, *14*(4), 347–375.

R Development Core Team (2006). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, *27*, 367–391.

Schnoebelen, T., & Kuperman, V. (2010). Using Amazon Mechanical Turk for linguistic research. *Psihologija*, *43*(4), 441–464.

Soames, S. (1982). How presuppositions are inherited: a solution to the projection problem. *Linguistic Inquiry*, *13*, 483–545.

Spector, B. (2003). Scalar implicatures: exhaustivity and Gricean reasoning. In B. ten Cate (Ed.) *Proceedings of the eigth ESSLLI student session*, (pp. 277–288). Vienna.

Spector, B. (2012). Bare numerals and scalar implicatures. To appear in: *Language and Linguistics Compass*.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*(1), 155–167.

Storto, G., & Tanenhaus, M. K. (2005). Are scalar implicatures computed online? In E. Maier, C. Bary, & J. Huitink (Eds.) *Proceedings of Sinn und Bedeutung 9*, (pp. 431–445). Nijmegen: Nijmegen Centre for Semantics.

van Rooij, R., & Schulz, K. (2004). Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information*, *13*, 491–519.

Zevakhina, N. (2012). Strength and similarity of scalar alternatives. In A. Aguilar Guevara, A. Chernilovskaya, & R. Nouwen (Eds.) *Proceedings of Sinn und Bedeutung 16*, (pp. 647–658). MIT Working Papers in Linguistics.

Zondervan, A. (2010). *Scalar implicatures or focus: an experimental approach*. Ph.D. thesis, Utrecht University.

Zweig, E. (2009). Number-neutral bare plurals and the multiplicity implicature. *Linguistics & Philosophy*, *32*(4), 353–407.

Appendix: Sentences used in Experiments 1 and 2

Notation: 'It ‖ The food (5) | salary (1) | solution (1) is adequate' means that in Experiment 1 the target sentence was 'It is adequate', while in Experiment 2 the target sentences were 'The food is adequate','The salary is adequate', and 'The solution is adequate', and that 'food', 'salary', and 'solution' were mentioned 5, 1, and 1 times, respectively, in the pretest where 10 participants were prompted for completions to the sentence 'The _____ is adequate but it is not good' (see the Materials section for Experiment 2).

Target sentences

• *adequate/good:* It ‖ The food (5) | salary (1) | solution (1) is adequate. • *allowed/obligatory:* It ‖ Copying (2) | Drinking (4) | Talking (2) is allowed. • *attractive/stunning:* She ‖ That ring (1) | This model (7) | The singer (2) is attractive. • *believe/know:* She believes it. The student (1) believes it will work out (1). The mother (3) believes it will happen (1). The teacher (6) believes it is true (1). • *big/enormous:* It ‖ That elephant (4) | The house (1) | That tree (1) is big. • *cheap/free:* It ‖ The water (2) | electricity (1) | food (5) is cheap. • *content/happy:* She ‖ This child (3) | The homemaker (1) | The musician (1) is content. • *cool/cold:* That ‖ The air (1) | weather (4) | room (1) is cool. • *dark/ black:* That ‖ That fabric (1) | The sky (3) | The shirt (1) is dark. • *difficult/impossible:* It ‖ The task (6) | journey (1) | problem (3) is difficult. • *dislike/loathe:* He dislikes it. The boy (1) dislikes broccoli (1). The teacher (2) dislikes fighting (1). The doctor (3) dislikes coffee (1). • *few/none:* He saw few of them. The biologist (1) saw few of the birds (2). The cop (1) saw few of the children (1). The observer (1) saw few of the stars (1). • *funny/hilarious:* It ‖ This joke (3) | The play (1) | This movie (7) is funny. • *good/excellent:* It ‖ The food (2) | That movie (2) | This sandwich (1) is good. • *good/perfect:* It ‖ The layout (1) | This solution (1) | That answer (1) is good. • *hard/unsolvable:* It ‖ That problem (6) | The issue (3) | The puzzle (5) is hard. • *hungry/starving:* He ‖ The boy (5) | dog (3) | elephant (1) is hungry. • *intelligent/ brilliant:* She ‖ The assistant (1) | That professor (2) | This student (3) is intelligent. • *like/love:* She likes it. The princess (2) likes dancing (1). The actress (1) likes the movie (1). The manager (1) likes spaghetti (1). • *low/depleted:* It ‖ The energy (2) | This battery (1) | The gas (5) is low. • *may/have to:* He may do it. The child (2) may eat an apple (1). The boy (3) may watch television (0). The dog (2) may sleep on the bed (1). • *may/will:* He may do it. This lawyer (1) may appear in person (0). The teacher (3) may come (2). The student (1) may pass (0). • *memorable/unforgettable:* It ‖ This party (2) | The view (1) | This movie (3) is memorable. • *old/ancient:* It ‖ That house (2) | mirror (1) | table (1) is old. • *palatable/delicious:* It ‖ The food (3) | That wine (2) | The dessert (1) is palatable. • *participate/win:* She ‖ The freshman (1) | runner (2) | skier (1) participated. • *possible/certain:* It ‖ Happiness (1) | Failing (2) | Success (2) is possible. • *pretty/beautiful:* She ‖ This model (5) | That lady (1) | The girl (4) is pretty. • *rare/extinct:* It ‖ That plant (3)

| This bird (2) | This fish (1) is rare. • *scarce/unavailable:* It || This recording (1) | resource (4) | mineral (2) is scarce. • *silly/ridiculous:* It || That song (3) | joke (6) | question (1) is silly. • *small/tiny:* It || The room (1) | The car (1) | This fish (2) is small. • *snug/tight:* It || The shirt (4) | That dress (2) | This glove (1) is snug. • *some/all:* He saw some of them. The bartender (1) saw some of the cars (2). The nurse (1) saw some of the signs (1). The mathematician (1) saw some of the issues (1). • *sometimes/always:* He is sometimes inside. The assistant (1) is sometimes angry (3). The director (1) is sometimes late (2). The doctor (2) is sometimes irritable (1). • *special/unique:* It || That dress (1) | That painting (1) | This necklace (1) is special. • *start/finish:* She || The athlete (1) | dancer (2) | runner (2) started. • *tired/exhausted:* He || The quarterback (1) | runner (1) | worker (3) is tired. *try/succeed:* He || The candidate (1) | athlete (1) | scientist (1) tried. • *ugly/hideous:* It || The wallpaper (2) | That sweater (1) | That painting (3) is ugly. • *unsettling/horrific:* It || The movie (6) | This picture (1) | The news (2) is unsettling. • *warm/hot:* That || The weather (5) | sand (1) | soup (3) is warm. • *wary/scared:* He || The dog (3) | victim (1) | rabbit (1) is wary.

Control sentences

• *clean/dirty:* That || The table is clean. • *dangerous/harmless:* It || The soldier is dangerous. • *drunk/sober:* He || The man is drunk. • *sleepy/rich:* He || The neighbor is sleepy. • *tall/single:* She || The gymnast is tall. • *ugly/old:* It || The doll is ugly. • *wide/narrow:* It || The street is wide.