Scalar diversity

Bob van Tiel
Radboud University Nijmegen

Emiel van Miltenburg
VU University Amsterdam

Natalia Zevakhina
National Research University
Higher School of Economics Moscow

Bart Geurts
Radboud University Nijmegen

Abstract

We present experimental evidence showing that there is considerable variation between the rates at which scalar expressions from different lexical scales give rise to upper-bounded construals. We investigated two factors that might explain the variation between scalar expressions: first, the availability of the lexical scales, which we measured on the basis of association strength, grammatical class, word frequencies, and semantic relatedness, and, second, the distinctness of the scalemates, which we operationalised on the basis of semantic distance and boundedness. It was found that only the second factor had a significant effect on the rates of scalar inferences.

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 lexical 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.

There is no uncontroversial definition of lexical scales. However, it is widely assumed that lexical scales contain expressions that are ordered in terms of informativeness and lexicalised to the same degree (e.g., Atlas & Levinson 1981, Gazdar 1979, Horn 1972). In this paper, we will confine our attention to scales that meet these minimal conditions. This means that we will not be concerned with ranked orderings or ad-hoc scales (e.g., Hirschberg 1991, Levinson 2000). All of the example scales in Table 1 count as lexical scales according to the traditional definition that we will adhere to.[1]

Category	**Examples**	
Adjectives	⟨intelligent, brilliant⟩	⟨difficult, impossible⟩
Adverbs	⟨sometimes, always⟩	⟨possibly, necessarily⟩
Connectives	⟨or, and⟩	
Determiners	⟨some, all⟩	⟨few, none⟩
Nouns	⟨mammal, dog⟩	⟨vehicle, car⟩
Verbs	⟨might, must⟩	⟨like, love⟩

Table 1: Sample scales for various grammatical categories.

The debate about scalar inferences has, for the most part, centered on the question of how these inferences come about. At least three answers to this question can be distinguished. The traditional view 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.

Several authors have proposed alternatives to this account. Levinson (2000), for example, stipulates that scalar terms are ambiguous between an interpretation with and without an upper bound; so 'some' is ambiguous between meaning 'at least some' and 'some but not all'. Chierchia, Fox, and Spector (2012) assume a similar

1. This overview does not include numerical expressions. Some authors have proposed that the upper bound associated with these expressions is caused by a scalar inference. This proposal has engendered a substantial theoretical and empirical literature, which runs to a large extent parallel to the literature about other lexical scales. See Spector (2013) for an overview.

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)	Banga et al. (2009)
	Geurts & Pouscoulous (2009)	Huang & Snedeker (2009)
	Clifton & Dube (2010)	Grodner et al. (2010)
	Barner et al. (2011)	Chemla & Spector (2011)
	Bott et al. (2012)	Geurts & van Tiel (2013)
	van Tiel (2014)	Degen & Tanenhaus (2014)
⟨or, and⟩	Noveck et al. (2002)	Storto & Tanenhaus (2005)
	Breheny et al. (2006)	Chevallier et al. (2008)
	Pijnacker et al. (2009)	Zondervan (2010)
	Chemla & Spector (2011)	
⟨might, must⟩	Noveck (2001)	
⟨start, finish⟩	Papafragou & Musolino (2003)	

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

ambiguity but at the syntactic rather than the lexical level. These authors postulate a silent syntactic operator whose meaning is similar to that of overt 'only'. Sentences with a scalar term are ambiguous between parses with and without that operator. If the operator is appended, (1) receives a reading that can be paraphrased as 'I ate only some of the cookies', thus excluding the upper bound.

A fair number of experiments have been conducted 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 interpretation, development, and processing of scalar inferences. A comparison with Table 1 makes it clear that several classes of scalar expressions, 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 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, this *uniformity assumption*, as we will call it, had not been questioned, but it was put to the test by Doran and colleagues (2009, 2012), 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 grammatical categories give rise to upper-bounded inferences. However, as we will argue in the following, there are a number of reasons for going over the same ground using a different task, which is what we did. Furthermore, we investigated a number of candidate explanations for the variability we observed.

1. Extant evidence for diversity

According to the uniformity assumption, observations about the behaviour of a particular lexical scale can typically be generalised to the whole family of lexical scales. Before Doran et al. put this assumption to the test, a number of experimental findings had already cast doubt on the view that all scalar expressions behave alike. For example, Noveck (2001) found that children and adults were more likely to interpret 'might' with an upper bound than 'some'. However, the experiments in which these scalar expressions were tested differed along a number of dimensions, thus precluding a straightforward comparison.

More direct evidence against the uniformity assumption comes from the interpretation of the existential quantifier in Dutch and French. This quantifier can be instantiated as 'enkele' or 'sommige' in Dutch, and as 'quelques' or 'certains' in French. Banga et al. (2009) found that 'sommige' licenses an upper-bounding inference more often than 'enkele'. Pouscoulous et al. (2007) found the same result for 'quelques' when compared to 'certains'. Moreover, a comparison between these studies shows that Dutch 'sommige' and 'enkele' were substantially more likely to be interpreted with an upper bound than their French counterpart 'certains'. These findings indicate that the likelihood of a scalar inference varies both within and between languages.

A similar conclusion can be drawn from 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 states of affairs where the scalar inference was false. For example, Bott and Noveck's (2004, experiment 3) participants rejected statements like those in (2) 59% of the time:

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

The main point transpiring from Geurts's survey is that, across the collated experiments, the mean rate of scalar inferences for 'or' was clearly lower than for 'some': 35% against 57%. This observation indicates that scalar inference rates are higher for 'some' than for 'or'.

There are also a number of developmental studies that have observed differences between lexical scales. Following up on Noveck (2001), Papafragou and Musolino

(2003) compared the rates of scalar inferences for three scales: ⟨some, all⟩, ⟨two, three⟩, and ⟨start, finish⟩. For adults, the rates of scalar inferences for these three scales were statistically indistinguishable, but children were significantly more likely to derive an upper bound for 'two' than for 'some' or 'start'. Similarly, Barner et al. (2011) found that children were significantly more likely to derive scalar inferences on the basis of an ad-hoc scale than on the basis of the lexical scale ⟨some, all⟩.

These preliminary observations aside, Doran et al. (2009, 2012) were the first to test the uniformity assumption in an integrated experimental design. In both of their studies, participants were presented 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 would you say Alex is doing financially?
Sam: He's comfortable.
FACT: Alex just bought four condos at Lake Point Tower, in downtown Chicago, where Oprah Winfrey lives.

Participants had to decide whether Sam's answers were true or false. The premiss was that if Sam's statement was deemed to be false, then participants must have derived a scalar inference.

One further manipulation introduced in Doran et al.'s first paper was that, in addition to the condition illustrated in (3) and (4), there were two other conditions: one in which Irene's question contained a scalar term that was stronger than the one used by 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. Would you say Alex is financially wealthy?
b. Would you say that Alex is poor, comfortable, or wealthy?

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 32% of the time, for sentences with adjectives, the rejection rate was only 17%. Scalar inferences were thus about twice as frequent for quantifiers as for adjectives. Secondly, Doran et al. found that only adjectival items were affected by the difference between the 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; otherwise, the neutral/contrastive distinction was inert.

Although Doran et al.'s findings provide convincing evidence against the uniformity assumption, there are a number of reasons for going over the same ground with a different experimental design and a finer-grained analysis. Firstly, Doran et al. adopted a rather coarse-grained categorisation of experimental items, grouping 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. Such a finer-grained analysis is also a prerequisite for determining what factors underlie the variable rates of scalar inferences.

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. A potential problem with this approach is that it is difficult to standardise the descriptions of the relevant facts. To illustrate, compare the fact descriptions in (3) and (4). A number of differences stand out. First, the fact description for 'comfortable' is more verbose than for 'most', which makes Sam's response seem almost like an ironic understatement in the case of 'comfortable'. Second, the fact description for 'most' contains the scalar expression 'entire' which is a possible scalemate of 'most'. This may have rendered the lexical scale for 'most' more available than for 'comfortable'. Such differences may have contributed to the results that Doran et al. found. We therefore repeated Doran et al.'s experiment using a different paradigm and a finer-grained analysis, and then considered a number of potential explanations for the observed variability.

2. New evidence for diversity

Instead of Doran et al.'s verification task, we decided to adopt an inference task, which has been widely used in the psychology of reasoning, and has occasionally been used in experimental studies on scalar inference (Chemla 2009, Geurts & Pouscoulous 2009). It has been shown that the inference paradigm yields higher rates of scalar inferences than the verification paradigm, but since we were primarily interested in relative frequencies of scalar inferences, that was no cause for concern.

Experiment 1

Participants

We posted surveys for 25 participants on Amazon's Mechanical Turk (mean age: 35; range: 21–63; 14 females).[2] 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.

Materials and procedure

Figure 1 shows an example of a critical item (the full list of materials is given in Appendix A). 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.

John says:

She is intelligent.

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

☐ Yes ☐ No

Figure 1: Sample item used in Experiment 1.

2. 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 (Buhrmester, Kwang, & Gosling 2011, Schnoebelen & Kuperman 2010, Sprouse 2011).

Materials comprised a selection of scales consisting of quantifiers (2 scales), adverbs (1), auxiliary verbs (2), main verbs (6), and adjectives (32). A complete list is given in Table 3. Our selection of scalar expressions was guided in part by examples discussed in the literature (e.g., Doran et al. 2009, Hirschberg 1991, Horn 1972). 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. In the final selection, we made sure to include 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 scales for which the opposite was true; we did this because we wanted to test the hypothesis that relative frequency has an effect on the rate at which a scalar inference is derived (Section 4.3).

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 A).[3]

Results and discussion

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. Control items were answered correctly on 94% of the trials. The results for the target trials are shown in Figure 2. It is evident from this graph that there was considerable variation among critical items, with positive responses ranging along a continuum from 4% (for seven adjective scales) to 100% (for ⟨cheap, free⟩ and ⟨sometimes, always⟩). The results of our first experiment thus disprove the uniformity assumption: different scalar expressions yield widely different rates of scalar inferences.

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. Though it is difficult

3. 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. The results of this pilot experiment correlated almost perfectly with the results from Experiment 1 ($r = .97$, $t(8) = 11.66$, $p < .01$). Apparently, the number of control items does not have a substantial effect on the contrasts between scales.

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

Table 3: List of scales used in the experiments reported in this paper. Legend: **SI** *= percentages of participants who 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);* **Lex** *= lexical class (*O *= open,* C *= closed) (Section 4.2);* **Freq** *= logarithm of the ratio between the frequency of the weaker scalar term and the frequency of the stronger scalar term (Section 4.3);* **LSA** *= semantic relatedness based on latent semantic analysis (Section 4.4);* **Dist** *= mean perceived semantic distance (Exp. 4);* **Bnd** *= boundedness (*+B *= bounded,* –B *= non-bounded) (Section 5.2).*

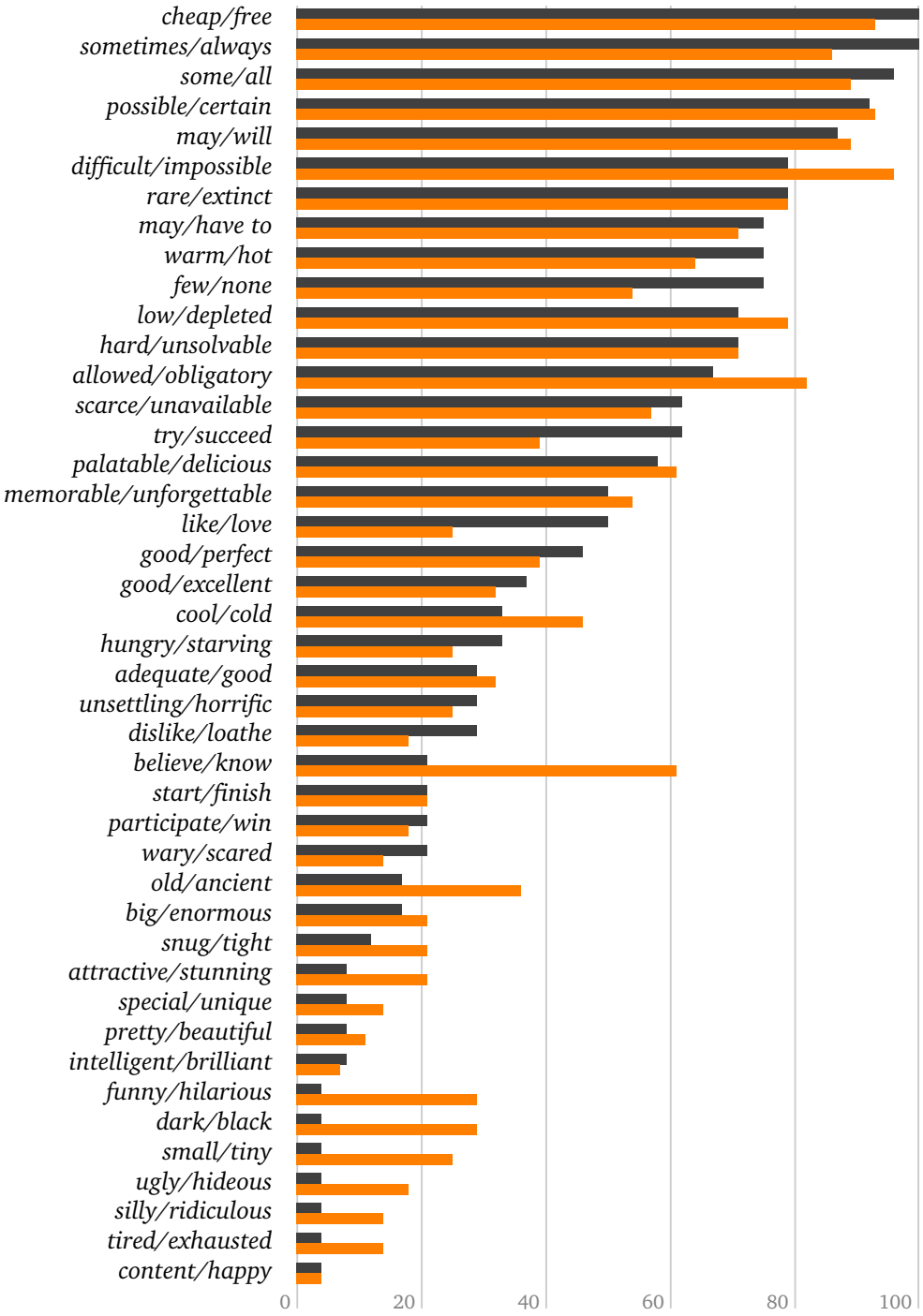

Figure 2: Percentages of positive responses in Experiment 1 (neutral content, dark grey) and Experiment 2 (non-neutral content, orange). The acceptance rates for entailments and unfounded inferences were 92% and 6%.

to see how this confusion could be responsible for the contrasts between scales, we thought it might be instructive to gauge the robustness of the results by replicating Experiment 1 with less neutral materials.

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. 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 and procedure

We tested the same scales as in Experiment 1, using the same procedure. However, in this case, the statements made 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:

(7) 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 because we wanted to avoid confusion about the meaning of the weaker scalar term. 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 (7a), 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. An example trial is given in Figure 3. 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. The target and control statements are listed in the Appendix A.

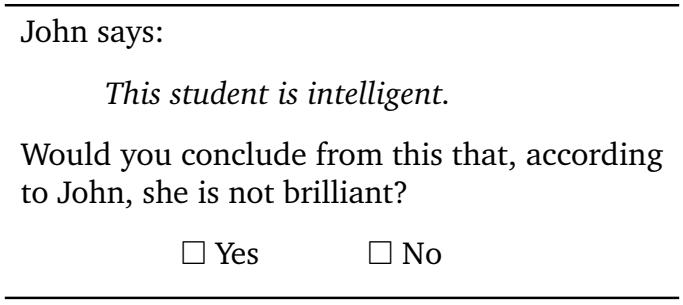

Figure 3: Sample item used in Experiment 2.

Results and discussion

One participant was excluded from the analysis for making mistakes in four control items. Four out of a total of 1500 answers were missing. Figure 2 shows the mean acceptance rates for each scale.

Paired chi-square tests showed that only two scales yielded different rates of scalar inferences in the two experiments, namely ⟨believe, know⟩, where the rate of positive responses increased from 20% to 60% ($\chi^2(1) = 7.42$, $p = .01$), and ⟨funny, hilarious⟩, where the rate of positive responses went from 4% to 30% ($\chi^2(1) = 4.05$, $p = .04$). Accordingly, the product-moment correlation between the proportions of positive answers for corresponding items in the two experiments was high ($r = .91$, $t(41) = 13.98$, $p < .01$). Overall, the rates of positive responses (42% versus 44%) did not differ significantly across the two experiments ($\chi^2(1) = 0.85$, $p = .37$). 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).

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 suggests that the general pattern of responses is robust to changes in the sentential context. 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 could hardly fluctuate more.

Before moving on, we first consider a potential methodological issue with the inference task. Consider the example trial in Figure 3. This trial asks participants if, according to the speaker, the student is 'not brilliant'. It has been observed that negated expressions sometimes cause an inference to the antonym. In other words, 'not brilliant' sometimes conveys a mitigated sense of dumbness (e.g., Fraenkel & Schul 2008, Horn 1989, Krifka 2007). Perhaps, then, the variable rates of scalar inferences that we observed in Experiments 1 and 2 are affected by the likelihood with which the negated scalemate licensed an inference to the antonym. According to this explanation, inferences to the antonym should occur more often with, for example, 'not exhausted' and 'not tight' than with 'not free' and 'not hot'.

There are, however, a number of reasons to assume that inferences to the antonym did not confound the general pattern of results. Firstly, the effect of inferences to the antonym might be preempted by the content of the speaker's statement. For example, participants might avoid interpreting 'not brilliant' as rather dumb because John just stated that she is intelligent. The question is much less trivial if the negated adjective receives its literal interpretation. Secondly, inferences to the antonym are especially robust if the negated expression contains a negative element itself (e.g., Horn 1989, Krifka 2007). We tested a number of such expressions: 'impossible', 'none', 'unsolvable', 'unavailable', and 'unforgettable'. However, all these expressions generated scalar inferences in more than 50% of the cases. Thirdly, Doran et al. (2009, 2012) compared scalar inference rates for quantifying expressions and gradable adjectives in a verification task. This paradigm does not involve negated expressions and is therefore not susceptible to the problem of inferences to the antonym. The relative proportions of scalar inferences for quantifying expressions and gradable adjectives in Doran et al.'s task (32% versus 17% negative responses) were the same as for scalar expressions from closed and open grammatical categories in Experiments 1 and 2 (76% versus 40% positive responses).

We conclude that the results of Experiments 1 and 2 provide a reliable indication of the likelihood with which different lexical scales license upper-bounding inferences. The variable rates of scalar inferences suggest that lexical scales differ in one or more aspects that are relevant for the computation of scalar inferences. In what follows, we discuss two such aspects: availability and distinctness. Afterwards, we measure the contribution of these factors to the rates of scalar inferences by operationalising them in a number of ways.

3. Explaining diversity

In order to compute a scalar inference, one has to assume that the speaker considered using a stronger scalemate of the scalar expression she used in his utterance. Otherwise it would be mistaken to infer from the speaker's utterance that she be-

lieves the stronger scalar expression is inappropriate. So perhaps the variable rates of scalar inferences are caused by differences in the *availability* of lexical scales.

Doran et al. (2009) provide some evidence to suggest that lexical scales are indeed available to different degrees. As discussed in Section 1, participants in their experiment were presented with stories in which Irene asked a question. In the neutral condition, Irene's question did not contain any scalar expressions; in the one-way contrastive condition, it mentioned a scalar expression that was stronger than the one used in Sam's answer; in the two-way contrastive condition, Irene's answer offered Sam three scalar expressions to choose from:

(8) a. How much cake did Gus eat at his sister's birthday party?
    b. Did Gus eat all of his sister's birthday cake?
    c. Did Gus eat some, most, or all of his sister's birthday cake.

It seems plausible that mentioning the scalemates of the scalar expression in Sam's answer makes the corresponding lexical scale more available and thus increases the likelihood of a scalar inference. In line with this prediction, Doran et al. observed higher rates of scalar inferences for adjectival scales in the two-way contrastive condition compared to the neutral and one-way contrastive conditions. No such effect, however, was found for quantificational scales. These observations can be construed as implying that quantificational scales are by default more available than adjectival scales. Explicit mentioning therefore has an effect on the rates of scalar inferences for adjectival but not quantificational scales.

Even if the lexical scale is available, a scalar inference can be preempted if the speaker used the weaker scalar term for a reason other than her believing that the utterance with the stronger scalar term is false. One such alternative reason is that the speaker is uncertain which scalar expression is appropriate. The likelihood that such a situation obtains will depend inter alia on the *distinctness* of the scale members, i.e., how easy it is to perceive the distinction between them. To illustrate, consider the scalar expressions 'some' and 'intelligent'. Intuitively, it is easier to establish if someone solved some or all of the problems than if a person is intelligent or brilliant. This difference in distinctness might explain why upper-bounding inferences were more frequent for 'some' than for 'intelligent'. More generally, the variable rates of scalar inferences may be attributable to differences in the distinctness of the scalar expressions on a scale.

In order to determine to what extent availability and distinctness can account for the variable rates of scalar inferences, we operationalised these notions in a number of ways. As measures of availability, we considered strength of association, grammatical class, word frequencies, and semantic relatedness. As measures of distinctness, we considered semantic distance and boundedness. In the following sections, we discuss these factors in greater detail.

4. Availability

4.1. Association strength

The most straightforward measure of the availability of a lexical scale is the strength of association between the scalar expression used in the speaker's utterance and its stronger scalemate. The greater the association strength, the more likely it is that the speaker considered using the stronger scale member. So perhaps the differential rates of scalar inferences can be explained in terms of differences in association strengths. To illustrate, consider the scalar expressions 'warm' and 'big'. The reason that scalar inferences were more frequent for 'warm' than for 'big' might be that the strength of association between 'warm' and 'hot' is much greater than between 'big' and 'enormous'. Thus we arrive at the following hypothesis:

> The availability of a lexical scale $\langle \alpha, \beta \rangle$ is an increasing function of the strength of association of β with α.

In order to test this hypothesis, we need to measure the strength of association between two scalar expressions. To this end, we conducted a modified cloze task. A standard cloze task, like the one we used to obtain materials for Experiment 2, 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.

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. 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

She is intelligent.

She is _____
She is _____
She is _____

Figure 4: Sample item used in Experiment 3 (–N condition).

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:[4]

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

4. 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 below, we paired the results for these statements with the results on the inference task for ⟨good, excellent⟩ and ⟨may, have to⟩, respectively. Changing this pairing did not have an effect on the results.

stronger scalar term we used in the inference tasks. However, this measure may be too strict because participants in the inference tasks might have computed a scalar inference based on a different stronger scalar term. 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 in the modified cloze task whether any stronger scalar term was mentioned. In this measure, we did not include scalar expressions that were stronger than the stronger scalar term we used in the inference tasks, such as 'perfect' for the ⟨adequate, good⟩ scale and 'freezing' for the ⟨cool, cold⟩ scale. After all, someone who infers from (9a) that, according to the speaker, it is not perfect does not necessarily infer that it is not good. Similarly for (9b): someone who infers that it is not freezing does not necessarily infer that it is not cold.

(9) a. It is adequate.
    b. That is cool.

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 4. 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%, $G^2(1) = 11.53$, $p < .001$). However, this effect interacted with the form of the instructions ($G^2(2) = 14.22$, $p = .001$): 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%, $G^2(1) = 7.22$, $p < .01$), and again there was an interaction with the neutral/non-neutral factor ($G^2(2) = 9.91$, $p < .01$): the effect reached significance for non-neutral statements only ($G^2(1) = 9.12$, $p < .005$).

A possible explanation for 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 the non-neutral condition, associations were constrained by the sentential context as well. To illustrate, compare the following sentences:

(10) a. That house is <u>old</u>.
     b. It is <u>old</u>.

	–N	+N
+S	25	29
–S	18	26

Strict coding

	–N	+N
+S	47	51
–S	40	46

Lenient coding

Table 4: 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).

Whereas in the case of (10a) participants might mention properties they associate with houses or old houses, (10b) is 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 46% of the trials. It was mentioned significantly more often with neutral than non-neutral sentences (49% vs. 44%, $G^2(1) = 6.41$, $p < .025$). As with the strict coding scheme, this effect interacted with the form of the instructions ($G^2(2) = 6.87$, $p < .05$): it only reached significance if the instructions did not contain a stronger scalar term ($G^2(1) = 5.01$, $p < .025$). Stronger scalar terms were mentioned significantly more often if the instructions contained a stronger scalar term than when they did not (49% vs. 43%, $G^2(1) = 9.57$, $p < .01$). There was an interaction with the neutral/non-neutral factor: the effect was only significant with non-neutral statements ($G^2(1) = 6.98$, $p < .01$).

Let us now examine the association hypothesis in light of the foregoing results. This and all of the following analyses were carried out using R, a programming language and environment for statistical computing (R Development Core Team 2006). In order to determine which factors are significant predictors of the rates of scalar inferences in Experiments 1 and 2, we used the `lme4` package (Bates & Maechler 2009) to construct a binomial mixed model with the responses in the inference tasks as dependent variable, and the measures with which we operationalised the notions of availability and distinctness as independent factors, including random slopes and intercepts for participants and items (Barr, Levy, Scheepers, & Tily 2013). The parameters of the mixed model are provided in Table 5 in Section 6.

The proportion of participants in Experiment 3 who mentioned a stronger scalemate was not a significant predictor of the rates of scalar inferences in the corresponding inference task ($\beta = 0.16$, $SE = 0.31$, $Z < 1$). The same conclusion holds for the strict analysis in which we counted the proportion of participants who mentioned the exact stronger scalemate that was used in the inference task

($\beta = 0.11$, $SE = 0.31$, $Z < 1$). Note that, for both measures, the direction of the effect is even opposite to what is predicted by the association hypothesis.

Therefore, whether or not a scalar inference is computed does not seem to depend on association strength, as operationalised in the modified cloze task. To illustrate, 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 scalar inference was only 16%; 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 of stronger scalar terms with their weaker scalemates.

It might be objected that the modified cloze task is a poor measure of association strength because participants who computed a scalar inference based on the target sentence might therefore not have mentioned a stronger scalar term. According to this explanation, participants were guided in part by the inferences that could be made on the basis of the target sentence. However, this prediction is incorrect, since antonyms were among the most frequently given answers: participants mentioned an antonym in 35% of the items. Apparently, participants were not constrained by the information conveyed by the target sentence. We thus conclude that association strengths do not have an effect on the rates of scalar inferences.

A more pressing issue is that the cloze task does not provide an absolute measure of the strength of association between two expressions. Even if the association strength between a scalar expression α and its stronger scalemate β is high, this might not be visible in the results of the cloze task because there are at least three expressions with which it is even more strongly associated. Conversely, even though the association strength between α and its stronger scalemate β is low, this might not be visible in the results of the cloze task because there are no other expressions with which it is more strongly associated. In order to address this concern, we implemented three other measures of availability. We leave open the question of how these measures relate to each other and to the underlying notion of availability.

4.2. Grammatical class

A first alternative measure of availability involves the distinction between open and closed grammatical classes. The domain of closed grammatical classes, like quantifiers and auxiliary verbs, is much smaller than that of open grammatical classes, like adjectives, adverbs, and main verbs. In consequence, the search space of alternatives is much smaller for closed grammatical classes than for open ones, and therefore it seems plausible to suppose that lexical scales are more available

when their elements are from a closed grammatical class than from an open one. The following hypothesis captures this explanation:

> The availability of a lexical scale $\langle \alpha, \beta \rangle$ is greater if α and β are from a closed grammatical class.

To test this hypothesis, we subdivided the scalar expressions into open and closed grammatical classes (Table 3). Although the average rate of scalar inferences was higher for scales from closed (76%) than open (40%) grammatical classes, the distinction between them did not have a significant effect on the rates of scalar inferences ($\beta = -0.47$, $SE = 0.47$, $Z = -1.00$, $p = .32$). One factor contributing to this nonsignificant result is that, in our experimental items, all closed-class scales were also bounded scales (but not the other way around). We discuss the distinction between bounded and non-bounded scales in Section 5.2.

4.3. Word frequencies

A third measure of availability is based on word frequencies. To see how these could have an effect, we compare the scales ⟨warm, hot⟩ and ⟨big, enormous⟩, which gave rise to scalar inferences 65% and 19% of the time, respectively. It might be that this discrepancy was caused by the fact that, whereas 'hot' is a quite common word that should be readily available to the speaker in a context in which she uttered 'warm', 'enormous' is rare relative to 'big', which might explain why the speaker did not use it even if, strictly speaking, it was more appropriate than 'big'. This explanation can be generalised and made more precise as follows:

> The availability of a lexical scale $\langle \alpha, \beta \rangle$ is an increasing function of 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. The logarithmised ratio of the frequencies of the scalemates did not have a significant effect on the rates of scalar inferences that we found in Experiments 1 and 2 ($\beta = -0.15$, $SE = 0.21$, $Z < 1$).

An alternative possibility is that it is not relative frequency, but rather the absolute 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, the frequencies did not have a significant effect on the results of Experiments 1 and 2 ($\beta = -0.14$, $SE = 0.24$, $Z < 1$).

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

4.4. Semantic relatedness

As a final test for the hypothesis that the variable rates of scalar inferences are caused by differences in the availability of the corresponding scale, we consider semantic relatedness. Words that are semantically related tend to occur in similar linguistic environments. To illustrate, 'warm' and 'hot' often co-occur with words like 'food', 'climate', 'water', and 'sand', whereas 'warm' and 'stunning' do not have such shared collocations. It has been demonstrated that words that tend to occur in the same environments also prime each other in word recognition tasks (Landauer, Foltz, & Laham 1998). It seems plausible to suppose, then, that semantic relatedness provides a good measure of availability:

> The availability of a lexical scale $\langle \alpha, \beta \rangle$ is an increasing function of the semantic relatedness of α and β.

A common measure of semantic relatedness is latent semantic analysis (Landauer & Dumais 1997). LSA constructs a matrix with words from a corpus as rows and columns. A row consists of binary values that represent whether the words in question occur in the same sentence; so words that co-occur in a sentence have a 1 in the same column. Words that are semantically related are expected to occur relatively often with the same words and thus have a lot of 1s in the same columns. Based on this matrix, LSA computes a value in the interval [0, 1] that denotes the semantic relatedness of different words. For example, the LSA value for 'warm/hot' is .51 as compared to .02 for 'warm/stunning'. Note that these LSA values do not reflect how often a pair of words co-occur, but rather how often they co-occur with the same words.

On the basis of Landauer, Foltz, and Laham's (1998) LSA implementation, we obtained relatedness 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 relatedness values, listed in Table 3, were used as an estimator of the results of Experiments 1 and 2. LSA values were not a significant predictor of the rates of scalar inferences ($\beta = 0.01$, $SE = 0.01$, $Z < 1$). We thus conclude that semantic

relatedness has no effect on the rates of scalar inferences that we observed in Experiments 1 and 2.

4.5. Conclusion

In order to compute a scalar inference, one has to assume that the speaker considered the corresponding lexical scale. Otherwise it would be mistaken to attribute her choice for a weaker scalar expression to the belief that the stronger scale member is inappropriate. Based on this observation, we hypothesised that the differential rates of scalar inferences in Experiments 1 and 2 were caused by differences in availability. In the foregoing sections, we operationalised the notion of availability by means of association strength, grammatical class, word frequencies, and semantic relatedness. But none of these measures made a significant contribution to the rates of scalar inferences. Availability thus plays at best a marginal role in shaping the results of Experiments 1 and 2.

It might be objected that the absence of a significant contribution of availability has a methodological cause. In our inference tasks, the question participants had to answer contained a scale member that was stronger than the one used in the target statement. One might suppose that this feature caused all lexical scales to be rendered available, thereby obviating the effect of intrinsic measures of availability likes the ones tested in the previous sections.

A number of observations speak against this explanation. First and foremost, recall that Doran et al. (2009) made a comparison between neutral, one-way contrastive, and two-way contrastive items. In the neutral condition, Irene's question did not contain scale members; in the one-way contrastive condition, it contained one scale member that was stronger than the one used in Sam's answer; and in the two-way contrastive condition, Irene, in effect, provided Sam with three scale members to choose from. The items in our inference tasks most closely resemble the items in Doran et al.'s one-way contrastive condition, since both involve a question that contains a scale member stronger than the one used in the target statement. Nevertheless, Doran et al. found no difference between the neutral and one-way contrastive items. This result provides strong evidence that mentioning a stronger scale member does not affect the availability of the lexical scale.

In addition, even if the question in the inference task made the lexical scale available to the participants, it does not follow that, according to these participants, it was also available to the speaker. After all, the question that mentions the stronger scalar expression was not presented to the speaker. In this respect, our inference tasks differ from Doran et al.'s one-way contrastive condition, in which the question that contains the stronger scalar expression was presented to the speaker character. So if mentioning a stronger scalar term affects the availability of lexical scales, this effect should be more pronounced in Doran et al.'s task than in our inference tasks.

The lack of an effect in Doran et al.'s task makes it unlikely that such an effect should have occurred in our inference tasks.

We conclude that availability plays a marginal role in determining the likelihood of a scalar inference. In the next section, we discuss a second possible factor: distinctness. If a scalar inference it computed, it has to be assumed that the speaker is able to determine which scalar expression is most appropriate. Therefore, if distinguishing between scalar expressions is difficult, it might be less likely that a scalar inference is derived. In the next section, we discuss two measures to operationalise the notion of distinctness: semantic distance and boundedness.

5. Distinctness

5.1. Semantic distance

The notion of semantic distance was inspired by an observation by Horn (1972, 90). Consider the following examples:

(11) 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.

An utterance of (11a) is more likely to implicate the negation of (11c) than the negation of (11b), since the negation of (11b) is logically stronger than the negation of (11c). So whenever a listener infers that the sentence with 'most' is false, she thereby also infers that the sentence with 'all' is false, but not vice versa. In more general terms, the likelihood of a scalar inference is an increasing function of the relative semantic distance between the scalar term used in the speaker's utterance and the stronger scalemate. See Zevakhina (2012) for an experimental analysis of how participants perceive such relative differences in semantic distance.

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:

> Given a lexical scale $\langle \alpha, \beta \rangle$, the distinctness of α and β is an increasing function of the semantic distance between these expressions.

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 likelihood of a scalar inference. 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.

(Note that the notion of semantic distance is not interdependent with the notion of semantic relatedness. It is possible for two expressions to be related but distant or unrelated but close. For example, 'warm' and 'cold' are related but distant.)

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. 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, which involved two statements which were synonymous or nearly so. These control items used the following pairs of words: 'enormous'/'immense', 'fantastic'/'sensational', 'gifted'/'talented', 'obvious/'clear', 'unbearable'/'intolerable', 'unexpected'/'unforeseen', and 'unpleasant'/'disagreeable'.

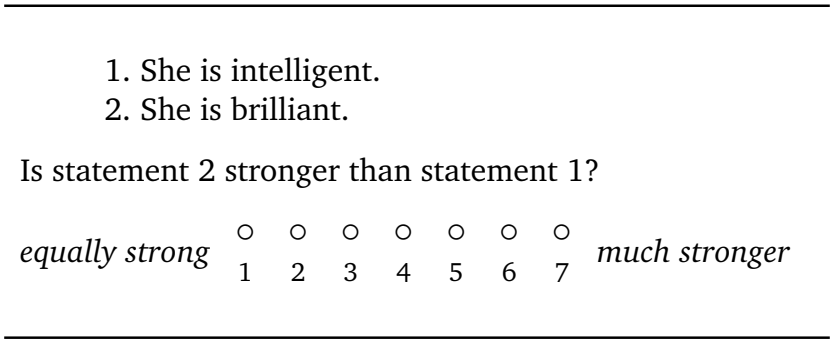

Figure 5: Sample item used in Experiment 4.

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 mean distance for the synonymous control items was 2.81. The 95% confidence interval around this mean was 2.53–3.09. There was only one lexical scale whose mean distance fell within that confidence interval: ⟨snug, tight⟩. This finding indicates that, except for this outlier, participants were able to perceive a difference in strength between scalemates.

The mean ratings on the distance task made a significant contribution to the rates of scalar inferences ($\beta = 0.65$, $SE = 0.27$, $Z = 2.36$, $p = .02$). This finding confirms the prediction made by the distance hypothesis. In Section 6, we discuss the variance explained by this and other factors.

5.2. Boundedness

A second measure of distinctness is more structural in nature. We have seen that rates of scalar inferences differ even within scalar expressions of the same grammatical class. For example, the percentages of positive responses for adjectival scales range

from 4% for ⟨content, happy⟩ to 95% for ⟨cheap, free⟩. However, there is an important difference between these two scales: in the case of 'cheap', but not in the case of 'content', the stronger scale member denotes an end point on the dimension over which the scalar terms quantify (Kennedy & McNally 2005, Rotstein & Winter 2004). We will refer to scales with such a terminal expression as *bounded*, as opposed to *non-bounded* scales like ⟨content, happy⟩. Note that boundedness depends on the semantics of the stronger scalar expression alone.

Scalar expressions on bounded scales can be distinguished on formal grounds alone: one scalar term denotes an interval and the other one an end point. By contrast, distinguishing scalar expressions on non-bounded scales requires inspecting the reach of the intervals denoted by both non-terminal expressions. It might therefore be hypothesised that scalar expressions on bounded scales are easier to distinguish than on non-bounded scales:

> Given a lexical scale $\langle \alpha, \beta \rangle$, the distinctness of α and β is greater if β is a terminal expression.

To test this hypothesis, we subdivided the lexical scales from Experiments 1 and 2 according to whether the stronger scalar expression denoted an end point, as can be seen in Table 3. It turned out that this classification subsumed the classification into open and closed grammatical classes. That is, all scalar expressions from closed grammatical classes occurred on bounded scales but not vice versa. This is not necessarily so: scales like ⟨some, most⟩ and ⟨sometimes, often⟩ are open even though they consist of elements from a closed grammatical class.

It was found that bounded scales indeed licensed higher rates of scalar inferences than non-bounded scales (62% versus 25%). Boundedness made a significant contribution to the rates of scalar inferences in Experiments 1 and 2 ($\beta = -1.87$, $SE = 0.40$, $Z = -4.72$, $p < .01$). The likelihood of a scalar inference is predicted in part by the distinction between bounded and non-bounded lexical scales. Section 6 discusses a measure of the variance explained by boundedness.[5]

5.3. Conclusion

If the distinction between scalar expressions is unclear, the speaker might choose to use a weaker expression because she is uncertain about whether the stronger expression is appropriate. Based on this observation, we hypothesised that the general pattern of results in Experiments 1 and 2 is shaped by the distinctness

5. One of the reviewers wondered to what extent the two measures of distinctness were correlated. The semantic distance between scalemates was perceived as greater when the stronger scalar term was a terminal expression than when it was not (5.28 versus 4.90) but this difference was only marginally significant ($t(41) = -1.71$, $p = .09$), which suggests that there was a small amount of overlap between these two factors.

of the scale members. In the previous sections, we operationalised the notion of distinctness by means of semantic distance and boundedness. Both of these measures turned out to have a significant effect on the rates of scalar inferences: the likelihood of a scalar inference increased with the semantic distance between scalar expressions, and scales with a terminal expression caused significantly higher rates of scalar inferences than scales without a terminal expression. We conclude that the likelihood of an upper-bounding inference is partly predicted by distinctness.

6. General discussion and conclusion

In recent years, neither the experimental nor the theoretical literature on scalar inferences has shown much concern for the diversity of scalar expressions, and by and large has confined its attention to less than a handful of items, notably 'some' and 'or'. Presumably, the tacit assumption has been that these are representative of the whole family of scalar terms. That assumption turns out to be mistaken: following up on studies by Doran et al. (2009, 2012), 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).

This was our main finding, but a large part of the foregoing discussion addressed the question of how the observed diversity can be accounted for. We considered two factors that might help to explain the variable rates of scalar inferences: availability and distinctness. Availability refers to how likely it is, according to the hearer, that the speaker considered stronger scalemates in the first place. Distinctness refers to how likely it is, according to the hearer, that the speaker considers the distinction between the weaker and the stronger scalar expression substantial enough that it is reasonable to assume that he should have used the latter if possible. In a series of analyses, we operationalised these factors in various ways.

We introduced two measures of distinctness, both of which made a significant contribution to the rates of scalar inferences:

i. Semantic distance
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 $\neg\varphi[\beta]$.

ii. Boundedness
Scalar expressions that inhabit a bounded scale, on which the stronger scalar term refers to an end point, were more likely to give rise to scalar inferences than their non-bounded counterparts. While bounded scales predominate in the upper half of the distribution in Figure 2, the lower half is populated mainly

by non-bounded scales. However, there is no strict dichotomy: inference rates were high for some non-bounded scales, too, and low for some of the bounded scales.

In contrast to these two measures of distinctness, none of our four measures of availability had a significant effect on the variable rates of scalar inferences:

i. Association strength
The probability that $\varphi[\alpha]$ gives rise to the inference that $\neg\varphi[\beta]$ might have 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 either hypothesis.

ii. Grammatical class
In their study, Doran et al. contrasted quantificational scales with adjectival scales. We included a similar subdivision between scalar expressions from open and closed grammatical classes. This distinction did not have an effect on the rates of scalar inferences.

iii. Word frequencies
The probability that $\varphi[\alpha]$ gives rise to the inference that $\neg\varphi[\beta]$, where β is a stronger scalemate of α, 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.

iv. Semantic relatedness
The probability that $\varphi[\alpha]$ gives rise to the inference that $\neg\varphi[\beta]$ might depend on how often α and β occur in similar linguistic environments. We determined the relatedness between expressions by means of latent semantic analysis (Landauer & Dumais 1997), but the outcome did not predict the rates of scalar inferences observed in Experiments 1 and 2.

In order to gauge how much variance was explained by each of the foregoing factors, we employed the measure of explained variance introduced by Nakagawa and Schielzeth (2012). The full mixed model, which included participants and items as random factors, and association strength, grammatical class, relative word frequencies, semantic relatedness, semantic distance, and boundedness as fixed factors, explained 52% of the variance in the results of Experiments 1 and 2 (Table 5). 22% of this variance was explained by the fixed factors and the remaining 30% by differences between items and participants. As for the independent factors, we found that none of our measures of availability explained more than 1% of the results. Distinctness turned out to be a more substantial factor, with semantic distance explaining 3% and boundedness explaining 10% of the results. Note that

these percentages do not sum to 22%, because some of the variance explained by a particular factor may be explained by another factor if the first factor is omitted from the model. For example, grammatical class explained a substantial part of the variance explained by boundedness in models where the latter factor was omitted.

Parameter	β	**SE**	**Z**	***p***	$\mathbf{R^2}$
(Intercept)	-2.80	1.73	-1.62	.104	–
Association strength	0.16	0.31	0.51	.611	.000
Grammatical class	-0.38	0.74	-0.52	.606	.001
Relative frequency	-0.15	0.21	-0.74	.461	.003
Semantic relatedness	0.01	0.01	0.93	.355	.006
Semantic distance	0.65	0.27	2.36	.018	.027
Boundedness	-1.87	0.40	-4.72	.000	.108

Table 5: Parameters of a mixed model with the results from Experiments 1 and 2 as dependent variable, the strengths of association based on the lenient coding scheme (Experiment 3), open or closed lexical class (Section 4.2), the logarithms of the ratio between the frequencies of scalemates (Section 4.3), the semantic relatedness between scalemates (Section 4.4), averages of the perceived semantic distance between scalemates (Section 5.1), and boundedness (Section 5.2) as independent variables, and random slopes and intercepts for participants and items.

To summarise, the full model explained roughly half of the observed variance; one fifth of the variance could be accounted for by factors we manipulated in our experiments, and half of that was due to boundedness. What could explain the remaining variance? One candidate factor that is often mentioned in the literature is that the likelihood 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 (12), but not in (13), should imply that Nigel has no more than fourteen children (examples taken from van Kuppevelt 1996):

(12) A: How many children does Nigel have?
B: Nigel has [fourteen]$_{\mathrm{F}}$ children.

(13) A: Who has fourteen children?
B: [Nigel]$_{\mathrm{F}}$ has fourteen children.

Since in our experiments no questions were asked, a possible explanation for the differential ratings of sentences with, e.g., 'warm' and 'big' is that participants tended

to contextualise these sentences in different ways, with 'warm' having a preference for a focus interpretation and 'big' having a preference a non-focus interpretation.

However, there are rather compelling reasons to doubt that this explanation is on the right track. In our experiments, scalar adjectives always occurred in predicate position, which is widely agreed to be focused by default (Ward & Birner 2004, 154). Furthermore, in Experiment 1, grammatical subjects were always pronominal, and pronouns rarely receive focus (*ibid.*, 158). To illustrate, it is obvious that, in the following examples, the adjectives are highly likely to be focused:

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

But whereas (14a) triggered scalar inferences in all cases, (14b) did so only 4% of the time. Although we cannot rule out the possibility that focus contributed to the rates of scalar inferences in Experiments 1 and 2, these observations suggest that it is not likely that focus was an important factor.

A second factor that might account for some of the remaining variance is the plausibility of the competence assumption. Starting with Soames (1982), scalar inference has often been treated as a two-step process, along the following lines (e.g., Geurts 2010, van Rooij & Schulz 2004, Sauerland 2004). 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 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.

Which brings us back to our initial question: How to explain the remaining variance in the data of Experiments 1 and 2? In the foregoing, we have looked at all the candidate factors we could think of. Almost none of these factors explained a substantial portion of the observed variance; the exception was boundedness,

and even its contribution was a mere 10%. In the absence of more successful candidates, we are forced to conclude that a major part of the observed variance was unsystematic. In Experiments 1 and 2, participants had to decide whether they would draw a scalar inference $\neg\varphi[\beta]$ from an utterance $\varphi[\alpha]$ that, save for the speaker's name, was not overtly contextualised. Making this decision requires an estimate of the likelihood that the speaker considered $\varphi[\beta]$ at least as relevant as $\varphi[\alpha]$. Our findings suggest that these estimates were by and large impervious to differences in word frequencies and various abstract semantic factors.

Perhaps it is not too surprising that this should be so. It is a well-established fact that speakers and hearers are alert to all manner of statistical patterns in language use (e.g., Seidenberg 1997), and therefore we might conjecture that language users keep track of the frequencies with which scalar expressions give rise to upper-bounded interpretations. If that is what underlies the remaining variance in Experiments 1 and 2, there is no reason to suppose that, e.g., the fact that sentences with 'silly' and 'tired' received the same rates of scalar inferences cannot be idiosyncratic.

It must be stressed that this line of reasoning is predicated on the absence of better explanations for our data, and is therefore highly tentative. However, if it is on the right track, it invites speculation about the processing of scalar expressions along the following lines. In the psychological literature, it is generally assumed that upper-bounded interpretations of scalars must be either defaults or due to an online inference (e.g., Bott & Noveck 2004, Breheny et al. 2006). But if it is true that, in our experiments, participants based their judgments on statistical patterns in their previous experience with scalar expressions, another view suggests itself. For it may be the case that, inside and outside the lab, hearers rely both on statistical regularities and on honest-to-Grice implicatures, employing the former to help them gauge the prior likelihood that an alternative expression will be relevant to the speaker, and the latter to derive their scalar inferences.

Even if an alternative is readily available, the speaker need not consider it sufficiently relevant to take it into account in his utterances. The concept of relevance is notoriously slippery, and it may not always be clear to the hearer whether or not a given alternative counts as sufficiently relevant or not. Whenever such quandaries arise, past experience may be brought to bear on the issue. If this picture is correct, the reason why young children are more cautious than adults in drawing scalar implicatures may be due, at least in part, to their more limited exposure to scalar expressions. The absence of a sufficient amount of past experience prevents them from associating utterances with their relevant alternatives and thus preempts a potential scalar inference.

In retrospect, it may have been a fortuitous incident that most of the experimental research on scalar inferences that has burgeoned since Bott and Noveck's (2004) landmark paper has been concerned with the interpretation of 'some'. Unlike many

other lexical scales, the connection between 'some' and 'all' is sufficiently strong to warrant the assumption that any cognitive effects associated with the interpretation of the weaker expression are due to the computation of the scalar inference rather than the association with its stronger scalemate. Nevertheless, it may be interesting to determine the role of statistical regularities on pragmatic inferencing by extending the scope of inquiry to other lexical scales as well.

Acknowledgements

We would like to thank Chris Cummins, Corien Bary, Ira Noveck, Judith Degen, Matthijs Westera, Paula Rubio-Fernández, Rob van der Sandt, Sammie Tarenskeen, Yaron McNabb, Ye Tian, and two anonymous reviewers for their comments and questions. This research was supported by a grant from the Netherlands Organization for Scientific Research (NWO), which is gratefully acknowledged.

BOB VAN TIEL
Department of Philosophy
Radboud University Nijmegen
bobvantiel@gmail.com

EMIEL VAN MILTENBURG
The Network Institute
VU University Amsterdam
emiel.van.miltenburg@vu.nl

NATALIA ZEVAKHINA
Faculty of Philology
National Research University
Higher School of Economics Moscow
nzevakhina@hse.ru

BART GEURTS
Department of Philosophy
Radboud University Nijmegen
brtgrts@gmail.com

References

Atlas, J. D., & Levinson, S. C. (1981). *It*-clefts, informativeness, and logical form. In P. Cole (Ed.) *Radical pragmatics*, (pp. 1–61). New York: Academic Press.

Banga, A., Heutinck, I., Berends, S. M., & Hendriks, P. (2009). Some implicatures reveal semantic differences. In *Linguistics in the Netherlands 2009*, (pp. 1–13). Amsterdam: John Benjamins.

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children's pragmatic inference. *Cognition*, *118*(1), 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`.

Bonnefon, J.-F., Feeney, A., & Villejoubert, G. (2009). When some is actually all: scalar inferences in face-threatening contexts. *Cognition*, *112*(2), 249–258.

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, *66*(1), 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*(3), 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*(3), 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.

Chemla, E. (2009). Universal implicatures and free choice effects: experimental data. *Semantics and Pragmatics*, *2*(2), 1–33.

Chemla, E., & Spector, B. (2011). Experimental evidence for embedded implicatures. *Journal of Semantics*, *28*(3), 359–400.

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.

Chierchia, G., Fox, D., & Spector, B. (2012). Scalar implicature as a grammatical phenomenon. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.) *An international handbook of natural language meaning, volume 3*, (pp. 2297–2332). Berlin: Mouton de Gruyter.

Clifton, C., & Dube, C. (2010). Embedded implicatures observed: a comment on Geurts and Pouscoulous (2009). *Semantics and Pragmatics*, *3*(7), 1–13.

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*(2), 128–133.

Degen, J., & Tanenhaus, M. K. (2014). Processing scalar implicature: a constraint-based approach. *Cognitive Science*.

Doran, R., Baker, R. E., McNabb, Y., Larson, M., & Ward, G. (2009). On the non-unified nature of scalar implicature: an empirical investigation. *International Review of Pragmatics*, *1*(2), 1–38.

Doran, R., Ward, G., Larson, M., McNabb, Y., & Baker, R. E. (2012). A novel paradigm for distinguishing between what is said and what is implicated. *Language*, *88*(1), 124–154.

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.

Fraenkel, T., & Schul, Y. (2008). The meaning of negated adjectives. *Intercultural Pragmatics*, *5*(4), 517–540.

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 and Pragmatics*, *2*(4), 1–34.

Geurts, B., & van Tiel, B. (2013). Embedded scalars. *Semantics and Pragmatics*, *6*(9), 1–37.

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*(1), 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. (1989). *A natural history of negation*. Chicago: Chicago University Press.

Huang, Y. T., & Snedeker, J. (2009). On-line interpretation of scalar quantifiers: insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*(3), 376–415.

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, *81*(2), 345–381.

Krifka, M. (2007). Negated antonyms: creating and filling the gap. In U. Sauerland, & P. Stateva (Eds.) *Presupposition and implicature in compositional semantics*, (pp. 163–177). Houndmills: Palgrave Macmillan.

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*(2-3), 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.

Nakagawa, S., & Schielzeth, H. (2012). A general and simple method for obtaining r^2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *4*(2), 133–142.

Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, *78*(2), 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*(2), 203–210.

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition*, *78*(3), 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*(4), 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.

Rotstein, C., & Winter, Y. (2004). Total adjectives vs. partial adjectives: scale structure and higher-order modification. *Natural Language Semantics*, *12*(3), 259–288.

Sauerland, U. (2004). Scalar implicatures in complex sentences. *Linguistics and Philosophy*, *27*(3), 367–391.

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

Seidenberg, M. S. (1997). Language acquisition and use: learning and applying probabilistic constraints. *Science*, *275*(5306), 1599–1603.

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

Spector, B. (2013). Bare numerals and scalar implicatures. *Language and Linguistics Compass*, *7*(5), 273–294.

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 Kuppevelt, J. (1996). Inferring from topics: scalar implicatures as topic-dependent inferences. *Linguistics and Philosophy*, *19*(4), 393–443.

van Rooij, R., & Schulz, K. (2004). Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information*, *13*(4), 491–519.

van Tiel, B. (2014). Embedded scalars and typicality. *Journal of Semantics*, *31*(2), 147–177.

Ward, G., & Birner, B. (2004). Information structure and non-canonical syntax. In L. R. Horn, & G. Ward (Eds.) *Handbook of Pragmatics*, (pp. 153–174). Malden, MA: Blackwell.

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.

Appendix A: 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 nurse (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.

Appendix B: Emotional valence

One of our reviewers suggested that emotional valence may have contributed to the variable rates of scalar inferences we found in Experiments 1 and 2. Bonnefon, Feeney, and Villejoubert (2009) demonstrated that the likelihood of a scalar inferences is influenced by considerations of politeness. Participants in their experiments were less likely to derive a

scalar inference if 'some' occurred in a face-threatening situation. For example, 'some' was less likely to be interpreted as 'some but not all' in (15b) compared to (15a):

(15) a. Some people loved your speech.
b. Some people hated your speech.

One explanation for this finding is that, in the case of (15b), a possible reason for the speaker to use 'some' instead of 'all' might be to avoid further damage to the listener's face. If that is indeed her motivation, it would be a mistake to conclude that the speaker believes the stronger alternative is false.

Based on this explanation, one might hypothesise that scalar expressions that have a negative connotation are less likely to be interpreted with an upper bound than scalar expressions with a positive connotation. To test this hypothesis, we presented 25 participants (mean age: 35; range: 23–72; 11 females), all of them U.S. residents and native speakers of English, on Mechanical Turk with the following instructions:

> Some words, like *fantastic* and *prosperous*, have positive associations. Other words, like *terrible* and *disappointing*, have negative associations.
>
> In the following, you will see a list of words. We ask you to indicate if these words are associated with positive or negative things by marking a value on a 7-point scale, where 1 means 'definitely negative', 7 means 'definitely positive', and 4 means 'neither negative nor positive'.

The list of words consisted of the stronger scalar terms used in Experiments 1 and 2. Including valence in the full model did not lead to a significant result ($\beta = -0.14$, $SE = 0.10$, $Z = -1.36$, $p = .175$). This finding suggests that emotional valence does not have a significant effect on the rates of scalar inferences.