Scales and scalarity: processing scalar inferences

Bob van Tiel

Elizabeth Pankratz

Chao Sun

Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)

Address for correspondence

Bob van Tiel

Leibniz-Zentrum Allgemeine Sprachwissenschaft (ZAS)

Schützenstrasse 18

10118 Berlin

Germany

bobvantiel@gmail.com

To appear in: *Journal of Memory and Language*

doi:10.1016/j.jml.2018.12.002

Abstract

The scalar word ‘some’ may be interpreted with an upper bound, i.e., as excluding ‘all’. Several studies have found that the computation of this *scalar inference* may be associated with a processing cost (e.g., Bott & Noveck, 2004; De Neys & Schaeken, 2007), which seems to argue in favour of theories according to which pragmatic inferencing is cognitively demanding (e.g., Sperber & Wilson, 1986). This argument holds on the premise that findings for ‘some’ can be generalised across the entire family of scalar words, which has been called into question by recent work highlighting the diversity within the class of scalar words (e.g., van Tiel, van Miltenburg, Zevakhina, & Geurts, 2016). In order to determine how generalisable the findings for ‘some’ are, we conducted three experiments in which we investigated the cognitive processing of seven scalar words that differ, inter alia, in their *scalarity*, i.e., whether they impose a lower (‘some’, ‘or’, ‘might’, ‘most’, ‘try’) or upper (‘low’, ‘scarce’) bound on their dimension. We find that the scalar inferences of the negatively scalar words ‘low’ and ‘scarce’ are not associated with a processing cost, unlike the scalar inferences of positively scalar words. We argue that the reported processing cost for scalar inferencing reflects increased cognitive demands associated with the processing of negative information.

Keywords: scalar inference; pragmatics; sentence processing; working memory; conversational implicature; language

Scales and scalarity: processing scalar inferences

A speaker who says (1) may imply that she did not eat all of the pie. If so, ‘some’, whose lower-bounded, literal meaning can be paraphrased as ‘at least some and possibly all’, receives an upper-bounded interpretation and thus comes to exclude ‘all’.

(1) I ate some of the pie.

The upper-bounded interpretation of ‘some’ is often explained as a *conversational implicature*, i.e., as an inference that can be worked out on the basis of the literal meaning of the utterance and the assumption that the speaker is cooperative (Grice, 1975). Thus, a speaker who says (1) could have been more informative—and, hence, cooperative—by producing the alternative in (2). Why didn’t she? Presumably because she did not eat all of the pie (e.g., Atlas & Levinson, 1981; Gazdar, 1979; Geurts, 2010; Horn, 1972; Soames, 1982).

(2) I ate all of the pie.

In order to describe this conversational implicature more concisely, it is often said that ‘some’ evokes the *lexical scale* <some, all>, and that an utterance containing ‘some’ may imply the negation of the corresponding sentence with ‘all’. For that reason, ‘some’ is called a *scalar word*, and the upper-bounding inference it may give rise to is referred to as a *scalar inference*.

There has been a prodigous amount of debate about the constituent properties of lexical scales (e.g., Gazdar, 1979; Hirschberg, 1991; Horn, 1972; Matsumoto, 1995). A first condition that is commonly identified is that words on a scale must be ordered in terms of informativeness. The notion of informativeness is usually operationalised in terms of logical strength, so that one word is more informative than another word if and only if it is logically stronger. Given this constraint, it follows that, e.g., <some, several> is not a well-formed lexical scale because ‘several’, as opposed to ‘all’, is not more informative than ‘some’.

At the same time, however, it is clear that informativeness alone is not a sufficient condition. For example, even though ‘some but not all’ is more informative than ‘some’, the scale <some, some but not all> is not well-formed. That is, (1) above clearly does not imply that (3) is false, i.e., it does not imply that the speaker ate all of the pie.

(3) I ate some but not all of the pie.

In order to account for this observation, Horn (1989, p. 234ff.) introduces a second constraint on lexical scales: their elements must have the same *scalarity*. The scalarity of a word refers to the type of bound it introduces on its dimension. ‘Some’ and ‘all’ are both *positively* scalar because they introduce a *lower* bound on, e.g., the amount of pie that was eaten. By contrast, ‘some but not all’ is non-scalar because it introduces both a lower and upper bound. Hence, <some, all> is a well-formed scale, but <some, some but not all> is not. Analogously, ‘not all’ and ‘none’ form a lexical scale because they are both *negatively* scalar in that they introduce an *upper* bound on, e.g., the amount of pie that was eaten. Hence, someone who utters (4) may imply that she did not eat none of the pie, i.e., that she ate some but not all of the pie.

(4) I did not eat all of the pie.

To further illustrate the notion of scalarity, suppose the pie is cut into six pieces. In that case, the lower-bounded meanings of the positively scalar words ‘some’ and ‘all’ can be visualised as in (5a), and the upper-bounded meanings of the negatively scalar words ‘not all’ and ‘none’ as in (5b). The corresponding scalar inferences of ‘some’ and ‘not all’ result in the negation of their higher-ranked scalemates ‘all’ and ‘none. Hence, in the case of ‘some’–and of positively scalar words more generally–the scalar inference is a negative proposition; in the case of ‘not all’–and of negatively scalar words more generally–the scalar inference is positive.

```
(5)   a.      some                    all
           ------[------------------------------[
           0    1    2    3    4    5    6
      b.   none                 not all
           ]------------------------------]------
           0    1    2    3    4    5    6
```

In summary, a lexical scale <α, β> is well-formed if the following two conditions are satisfied: (i) β is more informative, i.e., logically stronger, than α, and (ii) α and β are either both positively scalar or both negatively scalar.

There is a great variety of word pairs that satisfy these conditions. To illustrate, Table 1 shows a sample of positive and negative lexical scales consisting of words from different parts of speech (Doran, Baker, McNabb, Larson, & Ward, 2009; Gazdar, 1979; Hirschberg, 1991; Horn, 1972; van Tiel et al., 2016). In all of these cases, an utterance containing the first word on the scale may imply the negation of the corresponding sentence with the second word.

Category	*Examples*	
	Positively scalar	Negatively scalar
Adjectives	<intelligent, brilliant>	<silly, ridiculous>
	<warm, hot>	<cool, cold>
	<good, excellent>	<bad, horrible>
Adverbs	<possible, certain>	<improbable, impossible>
Connectives	<or, and>	
Determiners	<some, all>	<not all, none>
Nouns	<vehicle, car>	
Verbs	<like, love>	<dislike, loathe>
	<try, succeed>	<cut down, quit>

Table 1. Example positive and negative lexical scales consisting of words from different parts of speech.

Over the past two decades, a substantial number of studies have investigated the cognitive processing of scalar inferences. One of the more conspicuous results has been that, in certain experimental settings, the computation of scalar inferences is cognitively costly. Perhaps surprising, however, is that almost all of the studies that observed such a processing cost have been concerned with two positive lexical scales only: <some, all> and <or, and> (cf. van Tiel et al., 2016, p. 109). Our goal is to determine whether the findings that have been obtained for ‘some’ and ‘or’ can be generalised to the entire family of scalar words. More generally, our goal is to determine the source of the observed processing cost for scalar inferencing. Before we touch upon this issue, however, we provide a brief overview of the previous literature on the processing of scalar inferences.

Processing scalar inferences

In his book *Presumptive meanings*, Levinson (2000) argues that hearers automatically and effortlessly compute various types of pragmatic inferences, including scalar inferences. Thus, according to Levinson, someone who hears (6) immediately infers that the speaker did not eat all of the pie. In certain cases, this scalar inference may be cancelled, e.g., if the speaker continues with ‘In fact, I ate all of it’. According to Levinson, this process of cancellation should be cognitively effortful for the hearer, because it involves overturning a default interpretation.

(6) I ate some of the pie.

The predictions of *relevance theory* are diametrically opposite to what Levinson proposes (e.g., Sperber & Wilson, 1986, 1987). According to relevance theory, someone who hears (6) initially interprets the utterance literally, thus leaving it open whether or not the speaker ate all of the pie. If the hearer is not satisfied with the relevance of this interpretation, she may choose to make it more relevant by interpreting 'some' as 'some but not all'. According to relevance theory, this process of scalar inferencing involves a deeper processing of the utterance, which means that it should come at a cognitive cost.

In order to decide between these two theories, Bott and Noveck (2004) asked participants to indicate the truth value of sentences such as (7). These sentences are true when interpreted literally, but false if the scalar inference is computed and 'some' is interpreted as 'some but not all'. Thus, participants' truth judgements indicate whether or not they computed a scalar inference. For convenience, we will sometimes use the terms *literal* and *pragmatic* responses to refer respectively to 'true' and 'false' responses to underinformative sentences such as (7).

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

If hearers automatically interpret 'some' as 'some but not all', as Levinson argues, it seems natural to assume that literal responses should take at least as long as pragmatic responses. By contrast, if the computation of scalar inferences involves a deeper processing of the sentence, as relevance theory supposes, one may expect that pragmatic responses should take at least as long as literal responses.

In Bott and Noveck's Exp. 1, participants were explicitly instructed whether to respond literally or pragmatically to sentences like (7). Thus, in one condition, participants were told to interpret 'some' as 'at least some and possibly all'; in the other condition, as 'some but not all'. Bott and Noveck found that participants responded more slowly when they had to interpret 'some' as 'some but not all' than when they had to interpret 'some' as 'at least some and possibly all'. Crucially, this difference in response times was only present for underinformative sentences, such as (7), and not for control sentences, in which the truth value was independent of the way in which 'some' was interpreted, such as (8).

(8) a. Some mammals are dogs.
b. Some dogs are birds.

The delay in response times for pragmatic responses was confirmed in Bott and Noveck's Exp. 3. In this experiment, participants could provide their own intuitive truth judgements to underinformative sentences such as (7), rather than being trained to provide one type of response. Many participants in this experiment were ambivalent about the truth of these sentences, varying their responses across structurally similar trials. Comparing the response times of these ambivalent participants, Bott and Noveck found that it took them significantly longer to answer 'false' than 'true'. No such difference was found in the control condition, in which the truth value of the sentences was unambiguous, as in (8).

The computation of scalar inferences was thus associated with a delay in response times, which provides a strong argument against Levinson's proposal that the default interpretation of 'some' is upper-bounded (cf. Chemla & Bott, 2014; Cremers & Chemla, 2014; Noveck & Posada, 2003; Tomlinson Jr., Bailey, & Bott, 2013; van Tiel & Schaeken, 2016, for concurring evidence).

De Neys and Schaeken (2007) provide another telling piece of evidence against Levinson's account. As in Bott and Noveck's Exp. 3, participants in De Neys and Schaeken's experiment had to provide their intuitive truth judgements to sentences such as (7). While doing so, however, participants had to memorise patterns of dots on a 3x3 grid. In one condition, these dot patterns were simple, consisting of three dots in a horizontal or vertical line. In the other condition, the dot patterns were more complex, consisting of four dots scattered across the grid.

If the computation of scalar inferences involves deeper processing of the sentence, as relevance theory holds, participants should be less likely to respond pragmatically when their cognitive resources are taxed by having to memorise complex grids. By contrast, if the default interpretation of 'some' is two-sided, and it is the cancellation of this default interpretation that is cognitively effortful, as Levinson argues, participants should be *more* likely to respond pragmatically when they have to memorise complex grids.

In line with the relevance-theoretic prediction, De Neys and Schaeken found that participants were less likely to compute scalar inferences—i.e., to indicate that sentences such as (7) were false—when they had to memorise complex dot patterns than when they had to memorise simple dot patterns. This finding suggests that the computation of scalar inferences draws upon cognitive resources that were less available when participants had to memorise complex dot patterns (cf. Dieussaert, Verkerk, Gillard, & Schaeken, 2011; Marty & Chemla, 2013; Marty, Chemla, & Spector, 2013, for concurring evidence).

In what follows, we refer to Bott and Noveck's finding that participants' truth judgements to underinformative sentences are slower when 'some' is interpreted as 'some but not all' than when it is interpreted as 'at least some and possibly all' as the *B&N effect*, and to De Neys and Schaeken's finding that 'some' is more likely to be interpreted as 'at least some and possibly all' than as 'some but not all' when participants have to memorise complex dot patterns than simple ones as the *D&S effect*. The goal of this paper is to understand the source of these two effects for scales with different scalarity. Before turning to this issue, however, we briefly consider other studies that have investigated the processing of scalar inferences.

Data from other measures, such as reading times, event-related potentials (ERPs), and eye movements, paint a more complex picture about the presence or absence of a processing cost for scalar inferences, with some studies providing evidence in favour of relevance theory (e.g., Breheny, Katsos, & Williams, 2006; Chevallier, Bonnefond, Van der Henst, & Noveck, 2010; Noveck & Posada, 2003), and others supporting Levinson's defaultist approach (e.g., Barbet & Thierry, 2018; Nieuwland, Ditman, & Kuperberg, 2010; Politzer-Ahles & Husband, 2018). For reasons of space and relevance, we will only discuss the literature on eye movements in detail, referring the interested reader to the references above for further information on studies measuring reading times and ERPs.

A number of studies have recorded the eye movements of participants while they listened to sentences such as (9). The corresponding displays showed, among various other characters, a girl who had some but not all of the balls and a girl who had, e.g., all of the balloons. For convenience, we will refer to these characters as the *S-character* and the *A-character,* respectively. The question of interest is how quickly participants fixate on the S-character after hearing 'some'. If the computation of scalar inferences involves deeper processing of the sentence, as relevance theory holds, participants should initially distribute their attention over both the S-character and the A-character, converging on the S-character only at a later stage. By contrast, if 'some' is automatically interpreted as 'some but not all', as Levinson holds, participants should immediately fixate on the S-character.

(9) Click on the girl who has some of the balls.

Huang and Snedeker (2009, 2011) provide data that confirm the relevance-theoretic hypothesis; however, Grodner, Klein, Carbary, and Tanenhaus (2010) and Breheny, Ferguson, and Katsos (2013) provide data suggesting that participants immediately interpret ‘some’ as ‘some but not all’, thus confirming the defaultist prediction.

There are a number of differences between these studies that may have caused these discrepant results. Perhaps the most notable difference is that the eye-tracking studies of Huang and Snedeker also tested sentences containing number words, such as ‘two’ in (10), whereas the studies that confirmed Levinson’s defaultist account did not. These sentences with number words were also used to refer to S-characters, so that the characters in Huang and Snedeker’s studies could be referred to in different ways, whereas the characters in the studies that confirmed the defaultist predictions always had a unique description.

(10) Click on the girl who has two of the balls.

Huang and Snedeker (2018) argue that participants in the studies that supported Levinson’s defaultist account were already connecting characters with quantity words before they heard the sentence. So when participants heard ‘some’, they immediately concluded that the speaker had the S-character in mind rather than the A-character, which the speaker would have described using ‘all’. In Huang and Snedeker’s studies, the association between characters and descriptions was ambiguous, since the same character could be referred to with either a number word or a quantity word, which prevented participants from any kind of precoding. According to this explanation, then, the process of scalar inferencing is time-consuming unless the hearer already knows which quantity word goes with which character (see Degen & Tanenhaus, 2016, for an alternative explanation).

In the general discussion, we consider in more detail how the results from the eye-tracking literature relate to the results from studies that make use of the truth-value judgement paradigm. For the moment, however, our focus will be on the truth-value judgement studies, whose results seem to overwhelmingly confirm the relevance-theoretic view that the computation of the scalar inference of ‘some’ comes with a processing cost (but see Feeney, Scrafton, Duckworth, & Handley, 2004, for dissenting findings). This processing cost seems to provide a compelling argument against Levinson’s defaultist account. There are, however, a number of reasons to be sceptical of whether the B&N and D&S effects are truly indicative of a processing cost for scalar inferencing. Perhaps the prominent concern is that almost all of the current experimental evidence centers on only one scalar word, i.e., ‘some’, which is positively scalar, and that there are suggestive reasons for thinking that findings for ‘some’ may not straightforwardly generalise across all positively and negatively scalar words, as we will explain in the next section.

Broadening the scope

As noted in the introduction, the class of scalar words is extremely diverse, encompassing positively and negatively scalar words from various parts of speech (cf. Table 1). Research on the processing of scalar inferences, however, has focused primarily on the positive <some, all> scale (cf. van Tiel et al., 2016, Table 2). Various explanations may be given for the fixation on this particular scale. First, the meaning of ‘some’ is well-defined and independent of contextual factors, unlike the meanings of scalar words such as ‘probably’ and ‘few’. Second, it is relatively easy to construct underinformative sentences with ‘some’ using taxonomic information that all participants have access to. Third, ‘some’ is associated with a robust scalar inference, whereas the scalar inferences of other scalar words may not be quite as strong (van Tiel et al., 2016).

In addition, as one of our reviewers emphasised, it is not always necessary to test more than one scalar word to evaluate certain hypotheses. Thus, Bott and Noveck (2004) did not need to extend their purview to other scalar words to conclude that the predictions of Levinson’s defaultist account were falsified; the observation that the computation of the scalar inference of ‘some’ incurred a processing cost was sufficient. While Bott and Noveck—and, along the same lines, De Neys and Schaeken (2007)—thus provide strong evidence *against* the defaultist account, their data provide only limited inductive evidence in favour of relevance theory, unless of course the <some, all> scale is somehow representative for the entire family of lexical scales.

However, recent experimental work has called into question this uniformity hypothesis (e.g., Doran et al., 2009; Simons & Warren, 2018; van Tiel et al., 2016). Van Tiel and colleagues (2016) tested the derivation rate of 43 types of scalar inferences. To this end, they presented participants with a statement containing the weaker scalar word (e.g., ‘some’), and simply asked them if they would infer from that statement that the corresponding sentence with the stronger scalar word (e.g., ‘all’) is false. Van Tiel and colleagues found that the rates of scalar inferences varied dramatically across different lexical scales, ranging from close to 0% for scales such as <content, happy>, <tired, exhausted>, and <silly, ridiculous>, to almost 100% for scales such as <cheap, free>, <possible, certain>, and <some, all>.

Scalar words are thus extremely diverse in the rates at which they license scalar inferences. Of course, this should not be taken as evidence that scalar words also vary in the way they are processed. However, the differential rates of scalar inferences suggest we may observe a similar heterogeneity at the processing level. Moreover, there is further compelling evidence that different scalar words may be processed differently. As noted in the introduction, one of the few scales other than <some, all> that has been investigated in the processing literature is <or, and>.[1] However, the limited findings for the <or, and> scale that have been obtained are markedly less clear-cut than those for <some, all>: although, in line with relevance theory, Chevallier et al. (2008) provide data suggesting that the inference from ‘or’ to ‘not and’ is cognitively costly, Chevallier, Wilson, Happé, and Noveck (2010) fail to replicate the B&N effect for this scale.

A second, perhaps more compelling reason to doubt that the B&N and D&S effects generalise to all varieties of scalar inferences involves the negative <not all, none> scale. As noted in the introduction, sentences such as (11) may imply that the corresponding sentence with ‘none’ is false, i.e., that at least some dogs are insects. Two studies have investigated the processing of this scalar inference. While Cremers and Chemla (2014, Exp. 2) replicated the B&N effect for ‘not all’, both Cremers and Chemla (2014, Exp. 1) and Romoli and Schwarz (2015) observed a *reverse* B&N effect, with pragmatic responses to underinformative sentences with ‘not all’ being faster than literal responses.

(11) Not all dogs are insects.

At the very least, the limited findings for ‘or’ and ‘not all’ indicate that there are reasons to doubt that the processing of ‘some’ is exemplary for the entire family of scalar words. Specifically, the findings for ‘not all’ suggest that the processing of scalar inferences may be sensitive to the scalarity of scalar words. This potential lack of generality brings to the foreground a more fundamental question, namely, what is the source of the B&N and D&S effects? We discuss this question in the next section. Afterwards, we describe our own study,

[1] Here, we pass over the literature on the developmental trajectory of scalar inferences, which, despite a similar focus on <some, all> and <or, and>, has also investigated a number of other scales, including <start, finish> and <might, must>. Our motivation for passing over these studies in the present discussion is that they do not straightforwardly pertain to the debate between relevance theory and Levinson’s defaultist approach, tempting though it may be to connect developmental priority with ease of processing (cf. e.g., Barner, Brooks, & Bale, 2011; Noveck, 2001; Papafragou & Musolino, 2003; Pouscoulous, Noveck, Politzer, & Bastide, 2007).

in which we test various possible explanations for the B&N and D&S effects by extending the scope of inquiry to seven scalar words that differ, inter alia, in their scalarity.

Explaining the B&N and D&S effects

Why are ‘false’ responses to underinformative sentences with ‘some’ slower than ‘true’ responses, and why are participants who memorise complex dot patterns less likely to reject underinformative sentences with ‘some’ than participants who memorise simple dot patterns? The standard answer to these two questions is that the computation of scalar inferences is cognitively costly.

This explanation does not specify which aspect of the computation of scalar inferences is responsible for the processing cost. Initially, it was thought that the very process of pragmatic reasoning may be cognitively effortful (e.g., Bott & Noveck, 2004). However, recent findings suggest that other types of pragmatic inferences may be computed without any measurable processing cost (e.g., Chemla & Bott, 2014; van Tiel & Schaeken, 2016). In light of these results, it has been argued that the computation of scalar inferences is cognitively demanding because it involves constructing alternatives by substituting words in the uttered sentence with words from the lexicon (e.g., replacing ‘some’ with ‘all’, cf. Katzir, 2007). For our current purposes, however, this issue is immaterial. What is important is that, according to the standard explanation, the B&N and D&S effects are caused by an integral aspect of scalar inferencing. However, at least two alternative explanations may be given.

First, the two-sided interpretation of ‘some’ is semantically more complex than the lower-bounded interpretation. To elucidate, consider once again one of the underinformative sentences tested by Bott and Noveck (2004), repeated in (12).

(12) Some dogs are mammals.
⇝ Not all dogs are mammals.

On its literal interpretation, this sentence only places a lower bound on the number of dogs that are mammals. However, if the scalar inference is computed, and ‘some’ is effectively interpreted as ‘some but not all’, the sentence places both a lower and an upper bound on the number of dogs that are mammals. It has been suggested that this difference in complexity makes the two-sided interpretation of ‘some’ more difficult to process, which subsequently may have caused the B&N and D&S effects (e.g., Bott, Bailey, & Grodner, 2012; Geurts, 2010). In particular, this complexity-based explanation is often invoked to explain discrepancies between experiments measuring eye movements and truth-value judgement tasks (e.g., Degen & Tanenhaus, 2016; Grodner et al., 2010).

As one of our reviewers pointed out, this complexity-based account holds on the premise that people process the lower and upper bounds in serial rather than in parallel, which may be a tenuous assumption. Bott et al. (2012) provide further evidence against this complexity-based explanation based on data from a speed-accuracy trade-off task. However, the character of this task differed quite drastically from that of the simple truth-value judgement task employed by Bott and Noveck (2004), among others. Hence, we will, for the moment, suspend any misgivings about the semantic complexity explanation, and seriously conceive of the possibility that it may be correct.

Both the standard explanation and the semantic complexity explanation predict that, all else being equal, cognitive processing should be uniform across different scalar words and independent of their scalarity. In the previous section, however, we have seen that this assumption may be mistaken. Perhaps most prominently, some of the experimental evidence suggests that the inference from ‘not all’ to ‘some but not all’ is not associated with a processing cost and may even lead to faster response times.

One way of explaining the processing difference between ‘some’ and ‘not all’ centers on the observation, noted in the introduction, that ‘some’ is positively scalar, i.e., introduces a lower bound, whereas ‘not all’ is negatively scalar, i.e., introduces an upper bound. One of the consequences of this difference is that the scalar inference of ‘some’, i.e., ‘not all’, adds a *negative* proposition to the meaning of the utterance, whereas the scalar inference of ‘not all’, i.e., ‘not none’ or, equivalently, ‘some’, adds a *positive* proposition (cf. also Bott & Noveck, 2004, p. 454-455, for a similar argument).

There is a large body of evidence showing that the cognitive processing of negative information is difficult (e.g., Clark & Chase, 1972; Deschamps, Agmon, Loewenstein, & Grodzinsky, 2015; Geurts, Katsos, Cummins, Moons, & Noordman, 2010; Just & Carpenter, 1971; Moxey, 2006; Klatzky, Clark, & Macken, 1973; Rips, 1975). Various explanations have been given for why negative information should be difficult to process.

A first explanation revolves around the observation that negative sentences place constraints on the discourse that must be accommodated. In particular, negative sentences presuppose an expectation that their positive counterparts are true (e.g., Moxey, 2006; Moxey, Sanford, &

Dawydiak, 2001; Wason, 1965). In other words, negative sentences indicate a *shortfall* between the actual situation and an implicit or explicit expectation in the discourse. Thus, the negative scalar inference of ‘some’ in (12) may presuppose that there is an expectation in the discourse that all dogs are mammals. The accommodation of such presuppositions may make the processing of negative scalar inferences cognitively more demanding than the processing of positive scalar inferences, which do not carry such a presupposition.

A second explanation centers on the process of sentence verification. It has been argued that people verify negative sentences by first evaluating their positive counterparts and then reversing the truth value (e.g., Clark & Chase, 1972; Just & Carpenter, 1971; Rips, 1975). This verification process is more complex than for positive sentences, whose truth values are determined directly and independently of their negative counterparts. Thus, people may evaluate the scalar inference of ‘some’ in (12) by first determining whether all dogs are mammals and then reversing the truth value. The increased complexity of the verification procedure again may make the cognitive processing of negative scalar inferences more difficult than the processing of positive scalar inferences.

Hence, one may speculate that the scalar inferences of positively scalar words, but not negatively scalar words, are associated with a processing cost because they add a negative proposition to the meaning of the utterance. This scalarity-based explanation predicts no processing cost for the scalar inference of ‘not all’, in line with the data presented by Cremers and Chemla (2014) and Romoli and Schwarz (2015).

In order to test which of these explanations, if any, offers the most compelling account of the B&N and D&S effects, we redid two of Bott and Noveck’s truth-value judgement tasks, as well as De Neys and Schaeken’s memory load task, using sentences with seven scalar words that differ in their scalarity, as we explain presently.

Current study

In each of our three experiments, participants had to provide truth judgements in a sentence-picture verification task. We tested seven lexical scales: <low, empty> and <scarce, absent>, which are negatively scalar, and <or, and>, <might, must>, <some, all>, <most, all>, and <try, succeed>, which are positively scalar.[2] For each scale, we constructed three sentences

[2] Throughout the paper, we rely on an intuitive understanding of scalarity. There are various ways of formally confirming these intuitions (cf. Fauconnier, 1975; Horn, 1984; Kennedy & McNally, 2005). One such way focuses on the correlation between scalarity and *monotonicity*. Positively scalar words are *monotone increasing*,

containing the weaker scalar word, and, for each sentence, we constructed three types of pictures. In one type of picture, the sentence was unambiguously true, in one type of picture, the sentence was unambiguously false, and in one type of picture, the sentence was true when interpreted literally but false if the corresponding scalar inference was computed. The sentences with the first two picture types constitute the control condition; the sentences in the latter picture type form the target condition. Fig. 1 shows example sentences and pictures for each lexical scale.

Exp. 1 mirrored Bott and Noveck's (2004) Exp. 3. That is, participants simply provided truth judgements based on their own intuitions. We measured their response times to determine if 'false' responses were slower than 'true' responses in the target condition, when compared to the control condition. In a more exploratory analysis, we also investigated whether responses were correlated across scales, e.g., whether participants who computed the scalar inference for 'some' also tended to compute the scalar inference for 'low'.

Recent experimental work has observed that pragmatic inferences may prime each other (e.g., Bott & Chemla, 2016; Rees & Bott, 2018; van Tiel & Schaeken, 2016). Thus, Bott and Chemla (2016) found that participants were more likely to interpret 'four' as 'exactly four' rather than as 'at least four' when they had been primed with an upper-bounded interpretation of 'some'. Bott and Chemla argue that this priming effect suggests that the upper-bounded interpretations of both 'some' and 'four' share the same underlying mechanism. Although we did not carry out a full-fledged priming study, we may expect similar priming effects in that

i.e., they license inferences from sets to supersets. Hence, the inference patterns in (i) are valid, i.e., it is impossible that the premise is true but the conclusion is false, which shows that the scalar words 'some', 'or', 'might', 'most', and 'try' are positively scalar.

(i) a. Some people ate pepperoni pizza. ⇒ Some people ate pizza.
b. Bernie ate pepperoni pizza or pasta. ⇒ Bernie ate pizza or pasta.
c. Bernie might order pepperoni pizza. ⇒ Bernie might order pizza.
d. Most people ordered pepperoni pizza. ⇒ Most people ordered pizza.
e. Bernie tried to make pepperoni pizza. ⇒ Bernie tried to make pizza.

Conversely, negatively scalar words are *monotone decreasing*, i.e., they license inferences from sets to subsets. Hence, the inference patterns in (ii) are valid, which shows that the scalar words 'low' and 'scarce' are negatively scalar.

(ii) a. We are low on pizzas. ⇒ We are low on pepperoni pizzas.
b. Pizzas are scarce. ⇒ Pepperoni pizzas are scarce.

responses to different varieties of scalar inferences are correlated, assuming that they share the same underlying mechanism.

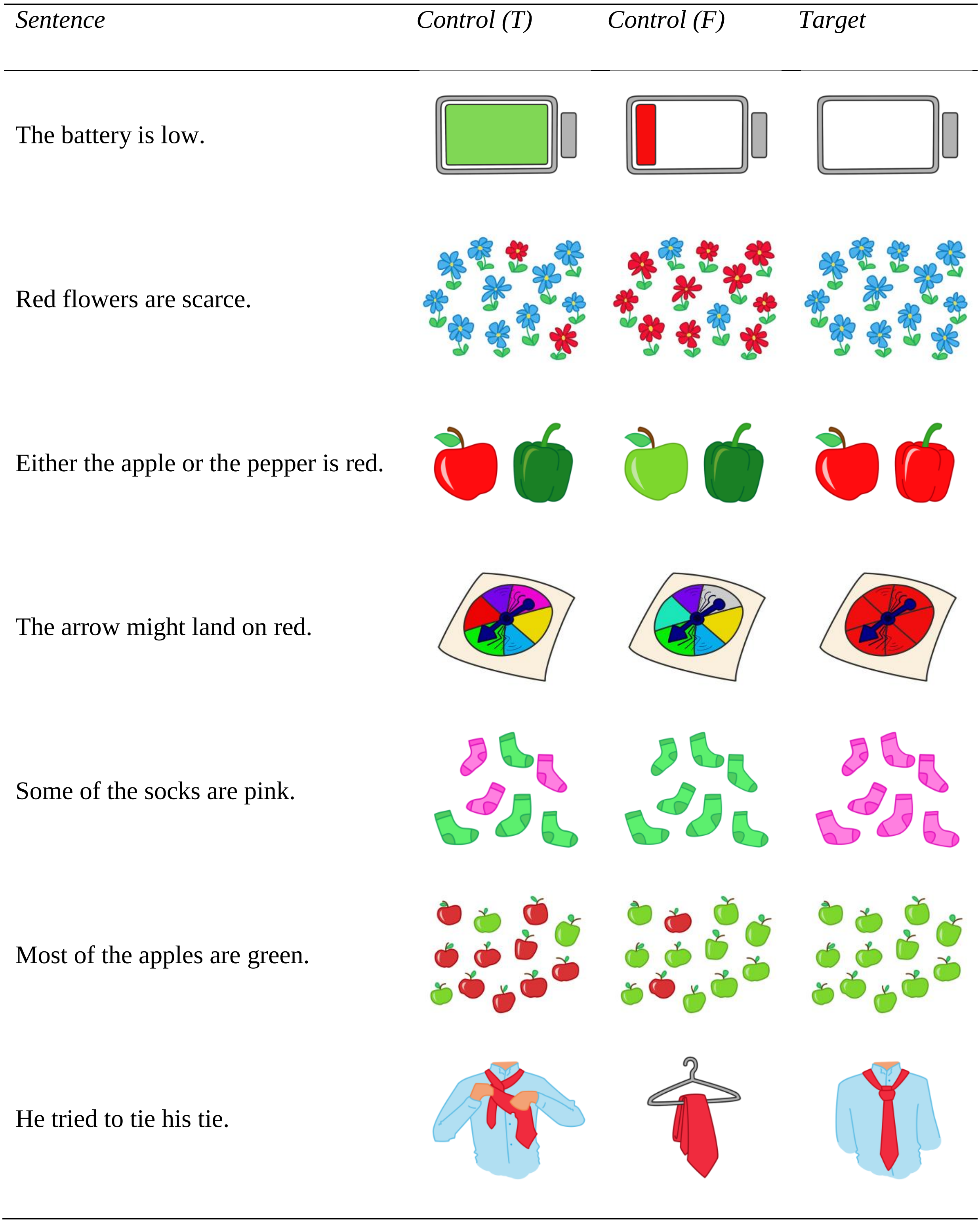
Sentence
Control (T)
Control (F)
Target
The battery is low.
Red flowers are scarce.
Either the apple or the pepper is red.
The arrow might land on red.
Some of the socks are pink.
Most of the apples are green.
He tried to tie his tie.

Figure 1. Sentences and example displays for each scalar term in Exp. 1.

Exp. 2 mirrored De Neys and Schaeken's (2007) study, and was essentially the same as Exp. 1. However, while providing their truth judgements, participants had to memorise 3x3 matrices containing black and white squares (cf. Bethell-Fox & Shepard, 1988; Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001). One group of participants had to memorise simple matrices, in which three black squares formed a horizontal line. Another group of participants had to memorise complex matrices, in which four black squares were scattered across the matrix. We determined whether participants who had to memorise the complex matrices were less likely to compute the seven varieties of scalar inferences than participants who had to memorise the simple matrices, as De Neys and Schaeken found for 'some'.

Exp. 3 mirrored Bott and Noveck's (2004) Exp. 1. Rather than providing truth judgements ad libitum, participants were instructed to interpret the seven scalar words either literally or with a scalar inference. After a sufficient amount of training, we measured participants' response times to determine whether the participants who were trained to interpret the scalar word literally were faster than the participants who were trained to compute the scalar inference, as Bott and Noveck found for 'some'.

In the next section, we lay out the predictions of the three explanations for the B&N and D&S effects that we introduced earlier. Afterwards, we describe the experiments in more detail.

Predictions

Earlier, we provided three possible explanations for the finding that, in truth-value judgement tasks, the realisation of the two-sided interpretation of 'some' is cognitively costly. According to the first explanation, the cognitive cost is due to the processing effort involved in computing scalar inferences. If this explanation is on the right track, we should find that all seven scalar words pattern with 'some' in that pragmatic responses take longer than literal responses, and in that pragmatic responses are less frequent when participants have to memorise complex matrices than when they have to memorise simple matrices.

The predictions of the second explanation are identical to the first. According to the semantic complexity explanation, two-sided meanings are intrinsically more difficult to evaluate than one-sided meanings. All of the scalar words that we investigated side with 'some' in that their literal meaning is one-sided, and that the scalar inference involves the addition of a second bound. Hence, the semantic complexity model also predicts that all scalar words pattern with 'some' in that the computation of their scalar inferences leads to a processing cost.

According to the scalarity-based explanation, only the scalar inferences of positively scalar words are cognitively effortful, since these introduce negative information into the meaning of the sentence. Like ‘some’, the scalar words ‘or’, ‘might’, ‘most’, and ‘try’ are all positively scalar. Thus, they place a lower bound on, respectively, the number of true disjuncts, the probability of an event, the number of entities that match the predicate, and the progress of an activity. By contrast, ‘low’ and ‘scarce’ are negatively scalar, since they stipulate an upper bound on, e.g., the energy level of a battery and the number of entities that match the predicate.

According to the scalarity-based explanation, then, pragmatic responses should be more effortful than literal responses for the positively scalar words ‘or’, ‘might’, ‘some’, ‘most’, and ‘try’, but there should be no difference in cognitive effort between these two types of responses for the negatively scalar words ‘low’ and ‘scarce’. Table 2 summarises the predictions of each of the three explanations for the B&N and D&S effects. In the next section, we describe the methods and results of the experiments in which we tested these different sets of predictions.

	low	scarce	or	might	some	most	try
Scalar inferencing	+	+	+	+	+	+	+
Semantic complexity	+	+	+	+	+	+	+
Scalarity	–	–	+	+	+	+	+

Table 2. For each of the three possible explanations, predictions whether or not (+ vs. –) each of the seven scalar words should be associated with the B&N and D&S effects. Note that the pragmatic inferencing and semantic complexity explanations make identical predictions.

Experiment 1: Responses and response times

Participants

50 participants (mean age: 37, standard deviation: 10, range: 21–65, 27 females) were drafted on Mechanical Turk and were paid $1.50 for their participation. Participants were asked to indicate their native language, but payment was not contingent on their response to this question. All participants indicated that they were native speakers of English.

Materials

The experiment tested seven scales: <low, empty>, <scarce, absent>, <or, and>, <may, must>, <might, will>, <some, all>, <most, all>, and <try, succeed>.[3] For each scale, we constructed one sentence with the weaker scalar term. These sentences were paired with three types of pictures. In one picture type, the sentence was unambiguously true (‘true’ control condition), in one picture type, it was unambiguously false (‘false’ control condition), and in one picture type, the truth value of the sentence depended on whether the corresponding scalar inference was computed (target condition). That is, the sentence was true if it was interpreted literally but false if the corresponding scalar inference was computed. Three slightly different tokens of each type of picture were created. Fig. 1 shows the seven sentences and example tokens of each picture type. The order of the items was completely randomised for each participant.

Procedure

Each trial started with the presentation of the sentence. Participants were instructed to press the space bar as soon as they had read and understood the sentence. Thereupon, the sentence disappeared and was replaced by a picture. Participants had to decide as quickly as possible whether or not the sentence was a good description of the depicted situation. They could register their decision by pressing either ‘1’ (good description) or ‘0’ (bad description) on their keyboard. Then, the picture disappeared and was replaced by the message ‘Press the space bar to continue’. Upon pressing the space bar, the next trial started.

Response times were recorded from situation onset to the point at which the ‘1’ or ‘0’ key was pressed. The entire experiment can be accessed online via http://spellout.net/ibexexps/muqtasp/si-div/experiment.html.

Data treatment

One participant was removed for making mistakes in more than 20% of the control items. 49 participants were thus included in the subsequent analyses. Their mean error rate on control items was 4.3%. In addition, due to a technical error, we had to remove data for the first item in each experiment (1.6% of the data). Finally, we removed items with a response time below

[3] Exp. 1 also tested an eighth scale, namely <may, have to>. However, data from this scale was not entered in the analyses, because participants made too many errors in the ‘true’ control condition (49% errors). For that reason, this scale was not tested in Exps. 2 or 3.

200 milliseconds or above 15 seconds, assuming that these correspond to accidental button presses or a lack of concentration on the task at hand (0.1% of the data).

Choice proportions

The percentages of 'true' responses for each scalar term and condition are summarised in Fig. 2. In the target condition of all scalar terms, participants were largely ambivalent about the truth value of the sentence. The percentages of 'true' responses in the target condition were: 'low': 71%, 'scarce': 54%, 'or': 47%, 'might': 50%, 'some': 38%, 'most': 60%, and 'try': 44%.

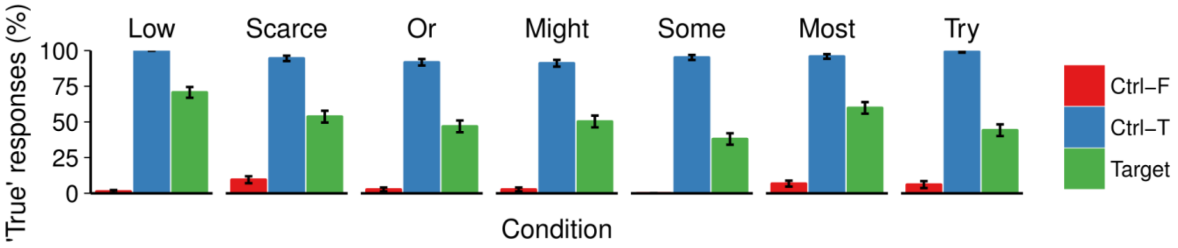

Figure 2. Percentage of 'true' responses for each scalar word and condition.

Consistency

In order to determine how consistently participants behaved across different scalar words, we determined the number of 'true' responses in the target condition for each participant and scalar word and analysed how well these correlated across scalar words. Table 3 shows the corresponding Kendall's rank correlation values, as estimated with the 'Kendall' package (McLeod, 2011).

In addition, we conducted a hierarchical agglomerative cluster analysis. For this analysis, we measured the Euclidean distance between the numbers of 'true' responses in the target condition of each scalar term for each participant. Thus, scalar words for which all participants gave a comparable number of 'true' responses in the target condition are close, whereas scalar words for which participants gave highly variable numbers of 'true' responses are distant. Using this distance metric, we clustered together the least distant scalar words stepwise, using Ward's method (Ward Jr., 1963). Fig. 3 shows a dendrogram visualising the outcome of this analysis.

	scarce	or	might	some	most	try
low	.30 *	.27 *	.26 ·	.24 ·	.27 *	.18
scarce		.26 *	.18	.30 *	.27 *	.28 *
or			.52 ***	.51 ***	.51 ***	.26 *
might				.64 ***	.61 ***	.14
some					.56 ***	.15
most						.11

Table 3. Kendall's τ_B rank-correlation between the number of 'true' responses in the target condition of each scalar term. Note: · indicates significance at the .10 level; * at the .05 level; ** at the .01 level; *** at the .001 level.

The results of the cluster analysis suggest a partitioning between two classes of scalar words: on the one hand, 'or', 'might', 'some', and 'most', and, on the other hand, 'low', 'scarce', and 'try'. Moreover, the results of both analyses indicate that the scalar terms in the former cluster behaved more similarly than the scalar terms in the latter cluster.

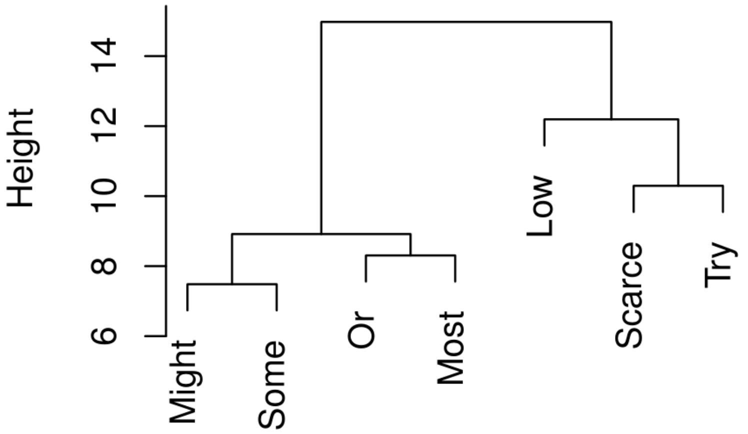

Figure 3. Dendrogram showing the results of a hierachical agglomerative cluster analysis using Ward's method based on the Euclidean distance between the numbers of literal responses in the target condition for each participant and scalar word.

Response times

Fig. 4 shows the mean logarithmised response times for each scalar word and condition. To analyse these response times, we constructed, for each scalar word, a linear mixed effects regression model predicting logarithmised response times on the basis of response ('true' or 'false'), condition (target or control), and their interaction, including random intercepts for participants. Random slopes for participants were dropped because only some of the models converged.

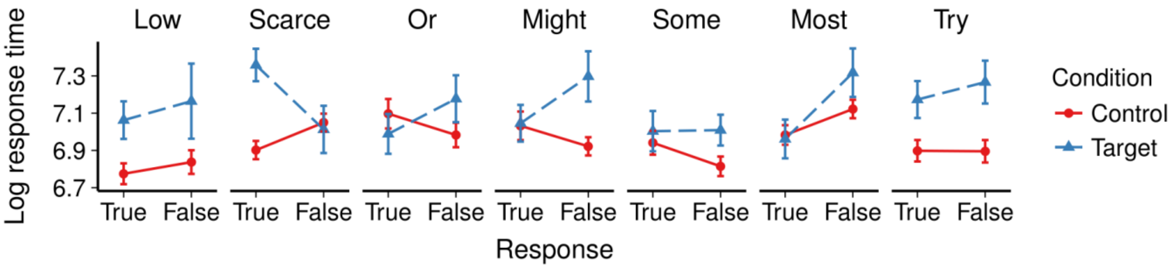

Figure 4. Mean log response times for each scalar term and condition in Exp. 1. Error bars represent within-participants standard errors (Cousineau, 2005).

The mixed models analyses, as well as all of the following analyses that will be reported, were conducted in R, a programming language and environment for statistical computing (R Development Core Team, 2006) using the ‘lme4’ package (Bates & Maechler, 2009). Degrees of freedom and corresponding *p*-values were estimated using the Satterthwaite procedure, as implemented in the ‘lmerTest’ package (Kuznetsova, Brockhoff, & Christensen, 2013).

The interaction between response and condition was significant for ‘or’ (β = -0.35, *SE* = 0.10, *t* = -3.49, *p* < .001), ‘might’ (β = -0.43, *SE* = 0.10, *t* = -4.49, *p* < .001), ‘some’ (β = -0.17, *SE* = 0.08, *t* = -2.05, *p* = .041), and ‘most’ (β = -0.23, *SE* = 0.08, *t* = -2.93, *p* = .004). The interaction was also significant for ‘scarce’, but going in the opposite direction (β = 0.49, *SE* = 0.08, *t* = 6.15, *p* < .001). The interaction was not significant for ‘low’ (β = -0.00, *SE* = 0.10, *t* < 1) or ‘try’ (β = -0.11, *SE* = 0.09, *t* = -1.30, *p* = .196).

In contrast with Bott and Noveck’s (2004) results, ‘false’ responses did not take longer than ‘true’ responses in the target condition of ‘some’, as shown by a linear mixed effects regression model predicting logarithmised response times in the target condition on the basis of response (‘true’ or ‘false’), including random intercepts for participants (β = 0.02, *SE* = 0.08, *t* < 1). A partial explanation for this discrepancy is that, in Bott and Noveck’s study, there was a pronounced response time bias facilitating ‘true’ responses. No such response bias was observed in our study. For ‘or’, ‘might’, and ‘most’, however, ‘false’ responses did take longer than ‘true’ responses in the target condition (all *p*’s < 0.025). Conversely, ‘true’ responses took longer than ‘false’ responses in the target condition of ‘scarce’ (β = 0.39, *SE* = 0.10, *t* = 3.87, *p* < .001). For ‘low’ and ‘try’, no difference in response times between ‘true’ and ‘false’ responses was observed (both *t*’s < 1).

To determine if the response time patterns differed across scalar words, we constructed, for each pair of scalar words, a linear mixed effects regression model predicting logarithmised response times on the basis of response (‘true’ or ‘false’), condition (target or control), scalar word, and all possible interactions. Again, these analyses only included random intercepts for participants due to issues of convergence. The significance of the three-way interactions between response, condition, and scalar word is shown in Table 4.

	scarce	or	might	some	most	try
low	4.75 ***	2.10 *	2.64 *	1.22	1.88 ·	< 1
scarce		6.82 ***	7.68 ***	5.86 ***	6.69 ***	5.12 ***
or			< 1	1.58	< 1	1.57
might				2.19 *	< 1	2.97 ***
some					1.37	< 1
most						1.49

Table 4. *t* and *p* values indicating whether the three-way interaction between response, condition, and scalar term had a significant effect on logarithmised response times. Note: · indicates significance at the .10 level; * at the .05 level; ** at the .01 level; *** at the .001 level.

The analyses of the interaction between condition and response suggest that the response time patterns fall into three categories. ‘Or’, ‘might’, ‘most’ showed a robust processing cost for ‘false’ responses in the target condition, compared to the control condition. The opposite pattern was observed for ‘scarce’. The remaining scalar terms fall in between these two groups, with ‘some’ tending towards the first group, and ‘low’ and ‘try’ not showing any interaction effect in either direction. However, the boundaries between these groups were not crisp. Thus, e.g., the response time patterns for ‘some’ and ‘most’ did not differ significantly from the response time patterns for ‘try’ and ‘low’.

In order to test whether the interaction between condition and response varied with the scalarity of the scalar word, we first coded the scalar words ‘low’ and ‘scarce’ as negatively scalar, and the other five scalar words as positively scalar. Afterwards, we constructed a linear mixed effects regression model predicting logarithmised response times based on scalarity (positive or negative), condition (target or control), response (‘true’ or ‘false’), and all of their interactions, including random intercepts for participants and scalar words. This model also included random slopes for the factors scalarity and response, which was the maximal converging model. In line with the scalarity-based explanation, there was a significant three-way interaction between scalarity, condition, and response (β = -0.48, *SE* = 0.07, *t* = -6.81, *p* < .001).

Experiment 2: Memory load

Participants

100 participants (mean age: 36, standard deviation: 11, range: 20–68, 48 females) were drafted on Mechanical Turk and were paid $2.00 for their participation. Participants were asked to indicate their native language, but payment was not contingent on their response to this question. Four participants were removed from the analyses for having a native language other than English.

Materials

The materials for Exp. 2 were mostly the same as the materials for Exp. 1, except that Exp. 2 did not test the scale <may, must>, which led to anomalous results in Exp. 1 (see Fn. 2). The order of the items was randomised for each participant.

Each trial started with the presentation of a pattern of black squares in a 3x3 matrix. Matrices in the low-load conditions contained three black squares on a horizontal line. These matrices thus contained *one-piece* patterns, i.e., all black squares were contiguous (Bethell-Fox & Shepard, 1988; Miyake et al., 2001). Matrices in the high-load condition contained four black squares. These matrices contained either *two-piece* or *three-piece* patterns, i.e., there were either two or three separate groups of contiguous black squares. Example matrices are shown in Fig. 5. The method of manipulating memory load was essentially the same as the method used by De Neys and Schaeken (2007). The only difference concerned the aesthetics of the matrices. In the study of De Neys and Schaeken, the matrices contained dot patterns; in our study, following the original study of Bethell-Fox and Shepard (1988), the matrix’s squares were completely filled.

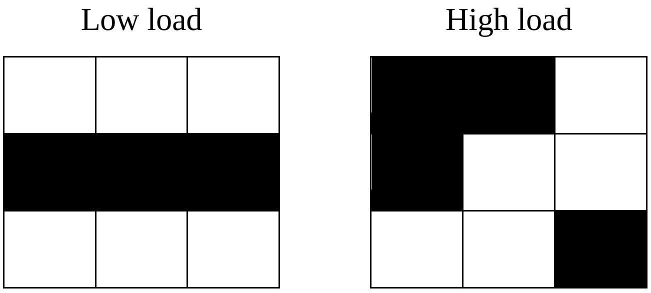

Figure 5. Examples of low-load and high-load matrices that participants had to memorise in Exp. 2.

Procedure

Each trial started with the presentation of a matrix, which appeared on screen for 850 milliseconds. Participants were instructed to memorise the pattern in these matrices. Afterwards, a sentence and a picture were presented in the middle of the screen. Participants had to decide as quickly as possible whether or not the sentence was a good description of the depicted situation. They could register their decision by pressing either ‘1’ (good description) or ‘2’ (bad description) on their keyboard. Once they had registered their decision, they saw an empty matrix and had to recreate the matrix that was presented at the start of the trial. To this end, participants could fill or unfill squares in the matrix by clicking on them. No feedback was given on their performance in this task. The high-load version of the experiment can be accessed via http://spellout.net/ibexexps/muqtasp/si-mem-h/experiment.html.

Data treatment

Results for one participant failed to register. Five participants were removed for making mistakes in more than 20% of the control items. Performance in the memory load task was measured by dividing the number of squares that were correctly characterised as filled or unfilled by the total number of squares. Five participants (all but one were in the high load condition) were removed from the analysis because their accuracy was lower than 75%. The mean accuracy of the remaining participants was 95% (low load: 96%, high load: 93%). 85 participants (low load: 45, high load: 40) were thus included in the subsequent analyses. Their mean error rate on control items was 3.8%.

We removed items with a response time below 200 milliseconds or above 15 seconds, assuming that these correspond to accidental button presses or a lack of concentration on the task at hand (0.3% of the data).

Memory load

In the subsequent analyses, we include the results of Exp. 1 as a baseline for the percentages of ‘true’ responses when there is no memory load. Fig. 6 shows the percentages of ‘true’ responses for each scalar term, condition, and memory load. In the control condition, performance was close to ceiling (all error rates < 10%). The only exception was the ‘true’ control condition for ‘or’ (no load: 8.3%, low load: 14.3%, high load: 16.7%). This finding is reminiscent of previous findings reported by Chevallier, Wilson, et al. (2010). About half of their adolescent participants rejected sentences with ‘or’ if only one of the disjuncts was satisfied. See Tieu et al. (2017) for an explanation of why some participants might reject sentences with ‘or’ if only one of the disjuncts is satisfied.

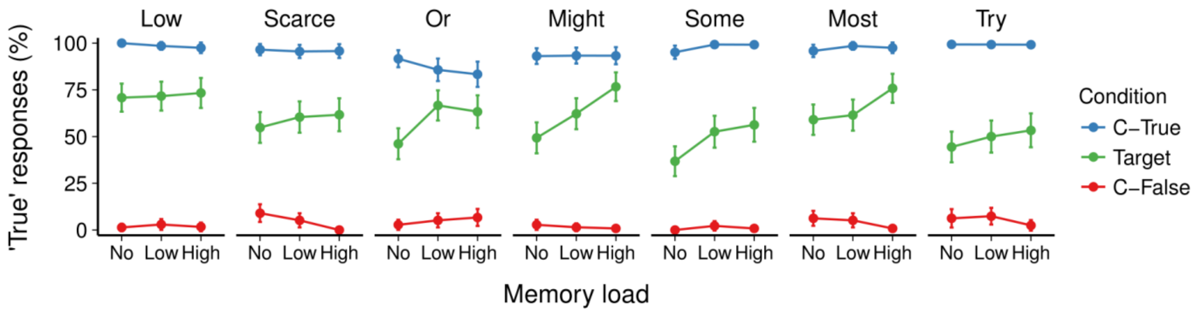

Figure 6. Percentage of ‘true’ responses for each scalar term, condition, and memory load. Error bars represent standard errors of the mean.

To analyse effects of memory load, we constructed, for the target condition of each scalar word, a generalised mixed effects logistic regression model predicting responses (‘true’ or ‘false’) on the basis of memory load (no load, low load, or high load), including random intercepts for participants. Memory load was included as an ordinal factor. We attempted to fit models with random slopes for participants, but these consistently failed to converge.

There were significant linear effects of memory load on the probability of ‘true’ responses for ‘or’ (β = 1.70, *SE* = 0.80, *Z* = 2.02, *p* = .043), ‘might’ (β = 12.10, *SE* = 1.65, *Z* = 7.32, *p* < .001), ‘some’ (β = 1.88, *SE* = 0.78, *Z* = 2.20, *p* = .028), and ‘most’ (β = -1.70, *SE* = 0.77, *Z* = 2.20, *p* = .027). There were no effects of memory load for the remaining scalar words (all *Z*’s < 1).

The effect of memory load for ‘some’ was mainly visible in the comparison between the no-load (i.e. Exp. 1) and low-load conditions, rather than between the low-load and high-load conditions, which would have been the expected result (De Neys & Schaeken, 2007; Marty et al., 2013). An explanation for this difference might lie in Dieussaert, Verkerk, Gillard, and Schaeken's (2011) observation that the effect of working memory load largely depends on the working memory capacity of participants. Specifically, the memory load effect only occurs for participants with a low working memory load capacity; it is harder for these participants to reach the pragmatic interpretation when their cognitive resources are taxed (Dieussaert et al., 2011, 2352). It may be the case that many participants in our study already experienced a maximal interference of memory load in the low-load condition. Thus, an even easier low-load condition may be necessary to recreate the D&S effect. In line with this hypothesis, more than half of the participants on Mechanical Turk have not enjoyed a university education (Pew Research Center, 2016). By contrast, both De Neys and Schaeken (2007) and Dieussaert et al. (2011, Exp. 2) tested university students. At the same time, however, Dieussaert et al. (2011, Exp. 1) tested students from a secondary school, but still observed the D&S effect, which speaks against our explanation, and suggests that other factors may modulate the presence or absence of the D&S effect. One speculative possibility is that participants are more focused on the task at hand when they are tested in real life rather than online.

To determine whether the effect of memory load differed across scalar words, we constructed, for the target condition of each pair of scalar words, a generalised mixed effects logistic regression model predicting response ('true' or 'false') on the basis of memory load (none, low, or high), scalar word, their interaction, and trial number. Again, these analyses only included random intercepts for participants due to convergence issues. The significance of the interactions between memory load and scalar term is provided in Table 5.

There was a significantly stronger effect of memory load on the probability of literal responses for 'or', 'might', 'some', and 'most' than for 'low'. In addition, the memory load effect was stronger for 'might' than for 'scarce' or 'try'. None of the other comparisons were significant.

Greater memory load thus led to fewer pragmatic responses for 'or', 'might', 'some', and 'most', although the effect was not always additive. The probability of pragmatic responses for 'low', 'scarce', and 'try' was independent of the degree of memory load that participants experienced. These results confirm the partitioning between these two groups of scalar words that we observed in Exp. 1.

	scarce	or	might	some	most	try
low	< 1	2.20 *	3.87 ***	2.28 *	2.22 *	< 1
scarce		1.54	3.21 **	1.88 ·	1.65 ·	< 1
or			1.95 ·	< 1	< 1	1.41
might				1.41	1.79 ·	3.21 **
some					< 1	1.60
most						1.22

Table 5. *Z* and *p* values indicating whether the interaction between scalar term and memory load had a significant effect on responses in the target condition for each pair of scalar terms. Note: · indicates significance at the .10 level; * at the .05 level; ** at the .01 level; *** at the .001 level.

In order to test whether the effect of memory load varied with the scalarity of the scalar word, we constructed a generalised mixed effects logistic regression model predicting response (literal or pragmatic) based on scalarity, memory load as an ordinal factor, and their interaction, including random intercepts for participants and scalar words. This model also included random slopes for the factor memory load, which was the maximal converging model. In line with the scalarity-based explanation, there was a significant interaction between memory load and scalarity ($\beta = 0.57$, $SE = 0.22$, $t = 2.60$, $p = .009$).

Experiment 3: Training

Participants

420 participants (mean age: 38, standard deviation: 12, range: 19–68, 197 females) were drafted on Mechanical Turk and were paid $0.60 for their participation. Participants were asked to indicate their native language, but payment was not contingent on their response to this question. 11 participants were removed from the analyses for having a native language other than English.

Materials

The experiment tested the same seven scalar words as Exps. 1 and 2. For each scalar word, we constructed three sentences, and for each sentence we created three types of pictures. As before, in one type of picture, the sentence was unambiguously true, in one type of picture, it was unambiguously false, and in one type of picture, the truth value of the sentence depended on whether the corresponding scalar inference was computed. Three slightly different tokens of each type of picture were created. One set of sentence and pictures was identical to the materials used in Exps. 1 and 2; the other two sets were created specially for Exp. 3. One set of sentence and pictures was exclusively used in the practice phase; the other two sets were exclusively used in the actual experiment.

Each participant only encountered sentences and pictures for one scalar word. There were thus 60 participants for each scalar term. The experiment itself always consisted of 18 trials: 12 control items and 6 target items. The actual experiment was preceded by 8 practice trials: 4 control items and 4 target items. The order of presentation was randomised for each participant.

Procedure

The procedure was essentially the same as for Exp. 1. However, half of the participants were instructed to interpret the scalar word literally, the other half was instructed to interpret it pragmatically. To illustrate, the instructions for ‘some’ said that participants should interpret this scalar word as either ‘some and possibly all’ (literal version) or ‘some but not all’ (pragmatic version).

These instructions were illustrated by means of an example. Afterwards, participants went through a practice phase consisting of 4 target items and 4 control items. These practice trials included feedback on participants’ performance and a reminder to interpret the scalar word in the appropriate way (either literally or pragmatically). Afterwards, the actual experiment started, in which participants did not receive any feedback on their performance. One version of the experiment can be accessed via http://spellout.net/ibexexps/muqtasp/train-or-log/experiment.html.

Data treatment

37 participants were removed for making mistakes in more than 20% of the control and target items. The accuracy of the remaining participants was 97% (target: 97%, control: 98%). The training was thus highly successful in steering participants towards either the literal or pragmatic interpretation of the scalar term. In total, 371 participants were included in the subsequent analysis.

We removed items with a response time below 200 milliseconds or above 15 seconds, assuming that these correspond to accidental button presses or a lack of concentration on the task at hand (0.1% of the data).

Response times

Fig. 7 shows the mean logarithmised response times for each scalar word and condition. To analyse these response times, we constructed, for each scalar word, a linear mixed effects regression model predicting logarithmised response times on the basis of response (‘true’ or ‘false’), condition (target or control), and their interaction, including random intercepts for participants. Random slopes for participants were dropped due to convergence issues.

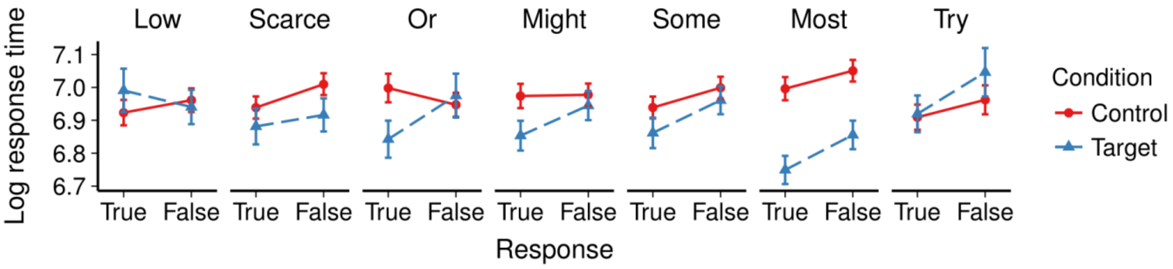

Figure 7. Mean response times for each scalar term and condition in Exp. 3. Error bars represent within-participants standard errors (Cousineau, 2005).

The interaction between response and condition was significant for ‘or’ (β = -0.26, *SE* = 0.06, *t* = -4.35, *p* < .001), ‘might’ (β = -0.14, *SE* = 0.05, *t* = -2.91, *p* = .004), ‘some’ (β = -0.10, *SE* = 0.05, *t* = -2.19, *p* = .029), ‘most’ (β = -0.11, *SE* = 0.05, *t* = -2.42, *p* = .016), and ‘try’ (β = -0.14, *SE* = 0.06, *t* = -2.24, *p* = .025). The interaction was also significant for ‘low’, but going in the opposite direction (β = 0.12, *SE* = 0.06, *t* = 2.12, *p* < .034). The interaction was not significant for ‘scarce’ (β = 0.02, *SE* = 0.05, *t* < 1).

The analyses of the interaction between condition and response largely confirm the results of Exp. 1. For ‘or’, ‘might’, ‘some’, and ‘most’, ‘false’ responses were significantly slower than ‘true’ responses in the target condition compared to the control condition. Unlike Exp. 1, the same response time pattern was observed for ‘try’. No such interaction was observed for ‘scarce’, and, in the case of ‘low’, the interaction went in the opposite direction.

In line with Bott and Noveck’s (2004) results, and unlike the results of Exp. 1, ‘false’ responses took longer than ‘true’ responses in the target condition of ‘some’, as shown by a linear mixed effects regression model predicting logarithmised response times in the target condition on the basis of response (‘true’ or ‘false’), including random intercepts for participants (*β* = -0.34, *SE* = 0.09, *t* = -3.65, *p* < .001). ‘False’ responses were also slower than ‘true’ responses in the target condition of ‘or’ (*β* = -0.26, *SE* = 0.09, *t* = -2.76, *p* = .008), ‘might’ (*β* = -0.22, *SE* = 0.09, *t* = -2.35, *p* = .023), ‘most’ (*β* = -0.32, *SE* = 0.09, *t* = -3.42, *p* = .001), and marginally for ‘try’ (*β* = -0.22, *SE* = 0.11, *t* = -1.97, *p* = .056). There were no significant effects of response for ‘low’ (*β* = 0.13, *SE* = 0.09, *t* = 1.48, *p* = .145) or ‘scarce’ (*β* = -0.05, *SE* = 0.10, *t* < 1).

To determine if the response time pattern differed across scalar words, we constructed, for each pair of scalar words, a linear mixed effects regression model predicting logarithmised response times on the basis of response (‘true’ or ‘false’), condition (target or control), scalar word, and all corresponding interactions. Again, these analyses only included random intercepts for participants due to issues of convergence. The significance of the three-way interactions between response, condition, and scalar word is shown in Table 6.

	scarce	or	might	some	most	try
low	1.30	4.60 ***	3.50 ***	3.08 **	3.19 **	3.05 **
scarce		3.60 ***	2.32 *	1.81 ·	1.97 *	2.00 *
or			1.48	2.04 *	1.86 ·	1.39
might				< 1	< 1	< 1
some					< 1	< 1
most						< 1

Table 6. *t* and *p* values indicating whether the three-way interaction between response, condition, and scalar term had a significant effect on logarithmised response times. Note: · indicates significance at the .10 level; * at the .05 level; ** at the .01 level; *** at the .001 level.

The results of Exp. 3 are largely concurrent with the results of Exps. 1 and 2. That is, the processing of the scalar inferences of ‘or’, ‘might’, ‘some’, and ‘most’ was cognitively demanding, as evidenced by slower response times for pragmatic responses than for literal responses. The converse was found for ‘low’, for which pragmatic responses were faster than literal responses. In the case of ‘scarce’, there was no difference in response times between pragmatic and literal responses. The most noteworthy difference between the results of Exp. 3

and the results of Exps. 1 and 2 was that, in Exp. 3, ‘try’ patterned with ‘or’, ‘might’, ‘some’, and ‘most’ in that pragmatic responses were slower than literal responses.

In order to test whether the interaction between condition and response varied with the scalarity of the scalar word, we constructed a linear mixed effects regression model predicting logarithmised response times based on scalarity (positive or negative), condition (target or control), response (‘true’ or ‘false’), and all of their interactions, including random intercepts for participants and scalar words. This model also included random slopes for the factors scalarity, condition, and response. In line with the scalarity-based explanation, there was a significant three-way interaction between scalarity, condition, and response (β = -0.22, *SE* = 0.05, *t* = -4.59, *p* < .001).

Accuracy

One of our reviewers pointed out that Bott and Noveck (2004, Exp. 1) also found an effect of scalar inferencing on participants’ accuracy. That is, participants who were trained to interpret ‘some’ as ‘some but not all’ made more errors on target items than participants who were trained to interpret ‘some’ literally. This finding again suggests that the computation of scalar inferences is cognitively effortful.

In order to determine if our data bear evidence of a similar pattern, we constructed a generalised mixed effects logistic regression model predicting responses (correct or incorrect) based on training (literal or pragmatic), including random intercepts for participants and scales. There was no significant effect of training (β = 0.29, *SE* = 0.57, *Z* < 1). Thus, in contrast with Bott and Noveck’s study, participants in the pragmatic training condition did not make more errors on target items than participants in the literal training condition.

In our study, participants performed at ceiling level across the target conditions of all scales (‘low’: 96% correct, ‘scarce’: 98%, ‘or’: 97%, ‘might’: 98%, ‘some’: 98%, ‘most’: 98%, and ‘try’: 95%) and in both training conditions (97% in both conditions), whereas participants in Bott and Noveck’s study regularly made errors (literal: ~90% correct, pragmatic: 60%). This discrepancy may be due to the fact that we only tested items with the weaker scalar word, whereas Bott and Noveck also tested items with the stronger scalemate, i.e., ‘all’. In their study, then, participants may have found it more challenging to focus on the trained meaning of the scalar word.

General discussion

Summary

Several studies using the truth-value judgement paradigm have found that the computation of the scalar inference from ‘some’ to ‘some but not all’ is cognitively demanding, in line with the relevance-theoretic view that the computation of scalar inferences involves deeper processing, and against Levinson’s (2000) idea that the default meaning of ‘some’ is two-sided. In order to determine whether the computation of scalar inferences is universally associated with a processing cost, and, more generally, what underlies this processing cost, we conducted three experiments to test the processing of seven lexical scales: <low, empty>, <scarce, absent>, <or, and>, <might, must>, <some, all>, <most, all>, and <try, succeed>. These scales differ, inter alia, in their scalarity, i.e., whether the words on the scale denote a lower bound (positively scalar) or an upper bound (negatively scalar) on their dimension. The scales <low, empty> and <scarce, absent> are negative; the other ones positive.

In Exp. 1, participants had to provide their intuitive truth judgements about sentences including the weaker scalar word in displays in which the sentences were literally true, but false if their scalar inferences were computed. We measured whether ‘false’ responses in these displays took longer than ‘true’ responses, as Bott and Noveck (2004) found for ‘some’. Exp. 2 was identical to Exp. 1, except that participants had to memorise simple or complex matrices while providing their truth judgements. We measured whether ‘false’ responses were less frequent when participants experienced greater cognitive load, as De Neys and Schaeken (2007) found for ‘some’. In Exp. 3, participants were trained whether to interpret the scalar words literally or with a scalar inference. As in Exp. 1, we measured whether ‘false’ responses in target displays took longer than ‘true’ responses, as Bott and Noveck found for ‘some’.

Table 7 provides an overview of the results of the three experiments. The scalar inferences of the positively scalar words ‘might’, ‘some’, ‘or’, and ‘most’ were consistently associated with a processing cost. The results for the negatively scalar words ‘low’ and ‘scarce’–as well as for the positively scalar word ‘try’–were more variable.

Towards an explanation

The standard explanation for the B&N and D&S effects is that they are due to the processing difficulty of computing scalar inferences. There has been some debate about which aspect of scalar inferencing takes time. Some authors have argued that the very process of pragmatic inferencing is time-consuming (e.g., Sperber & Wilson, 1986), while others have argued that it takes time to retrieve the stronger scalemate (e.g., ‘all’) from the lexicon (e.g., Chemla & Bott, 2014).

	low	scarce	or	might	some	most	try
Exp. 1	=	–	+	+	+	+	=
Exp. 2	=	=	+	+	+	+	=
Exp. 3	–	=	+	+	+	+	+

Table 7. Results of the three experiments. + indicates the presence of the B&N and D&S effects for that particular scale, – indicates an effect in the opposite direction, and = indicates no significant effect.

Our results for ‘low’, ‘scarce’, and, to a lesser extent, ‘try’ speak against both of these ideas, but are particularly problematic for the idea that the presence of a processing cost for scalar inferencing is exclusively determined by the ease of retrieving the stronger scalemate. ‘Or’, ‘might’, ‘some’, and ‘most’ are from closed grammatical classes, so, if anything, their stronger scalemates should be easier to retrieve than those of ‘low’, ‘scarce’, and ‘try’, which we found not to be associated with a processing cost (cf. also van Tiel et al., 2016, Exp. 3, for experimental data reinforcing this point).

A second, but in our study homologous, explanation for the B&N and D&S effects is that literal interpretations are intrinsically easier to process than two-sided interpretations. Again, this explanation predicts that all scalar words should give rise to the B&N and D&S effects—which is clearly contradicted by our results for ‘low’, ‘scarce’, and, to a lesser extent, ‘try’.

We also put forward a third explanation for the B&N and D&S effects. This scalarity-based explanation builds on the observation that positively scalar words, such as ‘some’ in (13), give rise to negative scalar inferences, whereas negatively scalar words, such as ‘scarce’ in (14), give rise to positive scalar inferences.

(13) Some of the flowers are red.
⇝ Not all of the flowers are red.

(14) Red flowers are scarce.
⇝ There are red flowers.

There is a large body of evidence showing that it is cognitively difficult to process negative information (e.g., Clark & Chase, 1972; Deschamps et al., 2015; Geurts, et al., 2010; Just & Carpenter, 1971; Klatzky et al., 1973; Moxey, 2006; Rips, 1975). Negative propositions have been argued to be difficult to process either because they presuppose an expectation that their positive counterparts are true (e.g., Moxey, 2006), or because they are evaluated indirectly by first determining the truth value of their positive counterparts (e.g., Clark & Chase, 1972). In either case, the scalarity-based explanation argues that, since the scalar inferences of positively scalar words introduce negative information, their computation should be cognitively effortful.

This proposal harmonises with the previously mentioned observation that there was no processing cost for the scalar inference of the negatively scalar term ‘not all’ (Cremers & Chemla, 2014, Exp. 1; Romoli & Schwarz, 2015). Further in line with this explanation, we found that the positively scalar words ‘or’, ‘might’, ‘some’, and ‘most’ gave rise to the B&N and D&S effects, whereas the negatively scalar words ‘low’ and ‘scarce’ did not. The scalarity-based explanation thus offers a much more compelling account of our results than the other two explanations that we considered.

The scalarity-based explanation makes a number of testable predictions. One prediction is that other lexical scales should pattern with the scales tested in our sample. That is, lexical scales with positively scalar items, such as <good, perfect>, <may, have to>, and <start, finish>, should also give rise to the B&N and D&S effects, as opposed to lexical scales with negatively scalar items, such as <cheap, free>, <few, none>, and <unlikely, impossible>. The main challenge in testing these scales will be to create suitable target and control items to which participants give the desired responses. In various pretests, we also attempted to create materials for testing the scales <may, have to>, <start, finish>, and <few, none>. However, participants almost universally accepted target items of the first two scales, and rejected target items for the third scale, which made the results for these lexical scales impossible to analyse.

The scales tested in our sample differ in a number of other respects that could, in principle, have influenced our results. First, the scalar words that we tested place thresholds on different types of dimensions: ‘low’ on the energy level of an object, ‘or’ on the number of true disjuncts, ‘might’ on the probability of an event, ‘some’, ‘most’, and ‘scarce’ on the number of individuals satisfying the predicate, and ‘try’ on the progress of an activity. Second, all of the negative scales in our sample were adjectival, and all of the positive scales were from other parts of speech. While we do not have specific hypotheses as to how these two factors could have affected our results, it may be reassuring to investigate pairs of scales that differ in their scalarity but that focus on the same dimension and come from the same parts of speech, such as <big, enormous> and <small, tiny>, <good excellent> and <bad, horrible>, or <like, love> and <dislike, hate>. If these pairs of scales behave in accordance with the scalarity-based explanation, this would rule out the possibility that our results are influenced by any other idiosyncrasies of the lexical scales included in our sample.

A second prediction made by the scalarity-based account follows from previous work showing that the processing cost for negative sentences disappears if these sentences are sufficiently contextualised. As noted previously, negative sentences presuppose an expectation that their positive counterparts are true (e.g., Moxey, 2006). Without adequate contextual support, this presupposition has to be accommodated, which makes the processing of these sentences more difficult. However, if this presupposition has already been satisfied in the prior discourse, the processing cost has been shown to disappear (e.g., Glenberg, Robertson, Jansen, & Johnson-Glenberg, 1999; Kaup, Yaxley, Madden, Zwaan, & Lüdtke, 2007; Tian, Breheny, & Ferguson, 2010).

In a similar way, it should be possible to ensure that the B&N and D&S effects disappear by making explicit an expectation that the positive counterpart to the scalar inference is true. Results from self-paced reading times experiments lend some support to this prediction. Thus, a number of experiments measured reading times for vignettes such as (15) and (16), taken from Politzer-Ahles and Husband (2013).

(15) Yousef asked Fatima whether any of the students had passed the test. Fatima said that some of them had. She added that the rest were planning to retake the class.

(16) Yousef asked Fatima whether all of the students had passed the test. Fatima said that some of them had. She added that the rest were planning to retake the class.

When deriving the scalar inference of ‘some’ in (15), participants have to accommodate the presupposition that there is an expectation that all of the students passed the test. In the case of (16), this presupposition has already been satisfied in the prior discourse. Correspondingly, in line with our prediction, a number of studies found increased reading times for (15) as compared to (16) (cf. Politzer-Ahles & Husband, 2013, Table 1, for a convenient overview). However, see also Bergen and Grodner (2012) and Breheny, Williams, and Katsos (2006) for data that speak against this prediction, and Politzer-Ahles and Fiorentino (2013) for an argument that the materials used in those studies were flawed.

The scalarity-based explanation thus offers a conceptually grounded and empirically testable account of our results. Nonetheless, it also faces a number of difficulties.

First, ‘try’, which is positively scalar, did not give rise to the B&N and D&S effects in Exps. 1 and 2, even though it did lead to a B&N effect in Exp. 3. Note, however, that even in Exps. 1 and 2, the results for ‘try’ were mostly statistically indistinguishable from the results for the other positively scalar words. The variation that we found in the case of ‘try’ may be due to the fact that it belongs to an open class, and the search in the lexicon for appropriate stronger alternatives—necessary to set the upper bound—may take longer than for the closed-class quantifiers, modals, and connectives. Further research is needed to disambiguate these possibilities.

Second, ‘scarce’ in Exp. 1 and ‘low’ in Exp. 3 gave rise to the *reverse* B&N effect, i.e., participants who computed the scalar inferences were faster than participants who interpreted the sentences literally. This finding is surprising on any approach that assumes pragmatic inferencing should be reflected in cognitive processing. Since scalar inferences add information to the literal meaning, their computation should not lead to faster response times. The only way of accounting for these reverse B&N effects is to invoke Levinson’s defaultist proposal, according to which hearers automatically compute scalar inferences. According to this proposal, a literal interpretation of sentences with scalar words requires the overturning of the default interpretation. However, we observed these reverse B&N effects in one experiment only, suggesting that they may be incidental rather than structural.

One may wonder how the scalarity-based explanation holds up in the face of data from other tasks. In the introduction, we discussed several eye-tracking studies on scalar inferencing. It has been reported in some of these studies, but not in others, that there is a delay in referential disambiguation based on pragmatically enriched ‘some’ relative to conditions where no pragmatic inferencing is involved.

If the scalarity-based explanation is correct, however, an alternative explanation is that the process of scalar inferencing itself is instantaneous, but that participants take longer to evaluate pragmatically enriched ‘some’ because of the added negative information. As discussed above, the processing of negative sentences tends to make salient their positive counterparts. In the case of ‘some’, then, the processing of the ‘not all’ inference would make the ‘all’ alternative salient. Perhaps this activation of the alternative hindered the rapid disambiguation of ‘some’. One way of testing this explanation would be to include trials with negatively scalar words, such as ‘not all’ in (17).

(17) Click on the girl who has not all of the balls.

If the scalarity-based explanation is on the right track, participants should immediately fixate on the girls who has some but not all of the balls in this condition.

Of course, it may turn out that future eye-tracking studies provide evidence that people immediately compute scalar inferences even in the absence of any precoding (cf. Degen & Tanenhaus, 2016). Such a finding would be problematic for the scalarity-based account. At the very least, it would mean that the scope of this explanation is restricted to truth-value judgement tasks.

Theoretical and methodological conclusions

There are two theories about the processing of scalar inferences. On the one hand, Levinson’s defaultist theory holds that scalar inferences are computed automatically and effortlessly; on the other hand, relevance theory posits that hearers initially interpret scalar words literally, and that the computation of scalar inferences requires deeper processing of the utterance. The current experimental record is equivocal about which of these two theories is correct. Results from eye-tracking studies, ERP studies, and self-paced reading time studies variously confirm the defaultist and relevance-theoretic predictions. However, data from truth-value judgement tasks consistently confirm the relevance-theoretic view that computing the scalar inference of ‘some’ is cognitively demanding, as the B&N and D&S effects reveal.

Our results indicate that it is questionable whether the B&N and D&S effects should be construed as evidence for the relevance-theoretic approach, since the source of these effects appears not to lie in the process of scalar inferencing itself, but rather in the fact that the scalar inference adds negative information to the meaning of the utterance. According to our results, the scalar inferences of negatively scalar words, such as ‘low’ and ‘scarce’, seem to be computed without any noticeable processing cost, in line with the defaultist predictions.

From a methodological point of view, our results have two important consequences. First, experiments on the processing of scalar inferences have hitherto been concerned almost exclusively with only two lexical scales: <some, all> and <or, and>. While the fixation on these two scales is understandable in light of the hypotheses that researchers addressed, this may have falsely given the impression that 'some' and 'or' are somehow representative for the entire family of scalar words. Our results show that this naive assumption is not warranted: there are marked differences in the ways in which different scalar words are processed. We recommend that future studies determine the generality of other processing findings by extending the scope of inquiry to a wider range of scalar words.

Second, most experimental research has focused on evaluating Levinson's defaultist account and its relevance-theoretic alternative. Our results show that this issue is not straightforwardly resolvable, because cognitive taxation is not a core property of scalar inferencing in general, but rather occurs only under certain conditions. A more suitable goal is to investigate under which circumstances the processing of scalar inferences leads to a processing cost, and explain why that should be so. While a number of studies have focused on this more modest goal (e.g., Degen & Tanenhaus, 2015; Gotzner & Benz, 2018; van Tiel, Kissine, & Noveck, 2018), there is still a pervasive assumption in the literature that the focus should be on confirming tout court either the defaultist or relevance-theoretic approach.

Conclusion

This contribution has sought to expand our understanding of the processing of scalar inferences by testing seven lexical scales belonging to various parts of speech and with different scalarity properties. We have shown that not all of these scales behave in the same way, indicating that the processing cost that has been suggested to be characteristic of scalar inferences (cf. Bott & Noveck, 2004; Chevallier et al., 2008, among others) may only hold under certain scalarity conditions: specifically, when the scalar term is positively scalar. Otherwise, the processing cost is not apparent. Our proposal is in harmony with previous research, which has often shown that the computation of the scalar inference 'some but not all' is an effortful step. However, it also cautions against the generalisation of results from only positive scales like <some, all> and <or, and> to all scalar inferences, since this blends out parts of the more nuanced picture. Future research should continue to explore the great diversity in scalar words in order to shed more light on the way scalar inferences are processed.

Acknowledgements

We would like to thank Richard Gerrig and the anonymous reviewers for valuable feedback on a previous version of this manuscript. This research was funded by the German Research Council (grant DFG FR 3482/2-1, KR951/14-1, SA 925/17-1) within SPP 1727 (Xprag.de), which is gratefully acknowledged.

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, NY: Academic Press.

Barbet, C., & Thierry, G. (2018). When *some* triggers a scalar inference out of the blue. An electrophysical study of a Stroop-like conflict elicited by single words. *Cognition*, *177*, 58-68. doi:10.1016/j.cognition.2018.03.013

Barner, D., Brooks, N., & Bale, A. (2011). Accessing the unsaid: the role of scalar alternatives in children’s pragmatic inference. *Cognition, 118,* 84–93. doi:10.1016/j.cognition.2010.10.010

Bates, D., & Maechler, M. (2009). lme4: linear mixed-effects models using S4 classes [R package]. Retrieved from http://cran.r-project.org/package=lme4

Bethell-Fox, C. E., & Shepard, R. N. (1988). Mental rotation: effects of stimulus complexity and familiarity. *Journal of Experimental Psychology: Human Perception and Performance, 14,* 12–23. doi:10.1037/0096-1523.14.1.12

Bergen, L. & Grodner, D. (2012). Speaker knowledge influences the comprehension of pragmatic inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 1450–1460. doi:10.1037/a0027850

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, *66*, 123–142. doi:10.1016/j.jml.2011.09.005

Bott, L., & Chemla, E. (2016). Shared and distinct mechanisms in deriving linguistic enrichment. *Journal of Memory and Language*, *91*, 117–140. doi:10.1016/j.jml.2016.04.004

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language, 51*, 437–457. doi:10.1016/j.jml.2004.05.006

Breheny, R., Ferguson, H. J., & Katsos, N. (2013). Taking the epistemic step: toward a model of on-line access to conversational implicatures. *Cognition*, *126*, 423–440. doi:10.1016/j.cognition.2012.11.012

Breheny, R., Katsos, N., & Williams, J. (2006). Are generalized scalar implicatures generated by default? An online investigation into the role of context in generating pragmatic inferences. *Cognition*, *100*, 434–463. doi:10.1016/j.cognition.2005.07.003

Chemla, E., & Bott, L. (2014). Processing inferences at the semantics/pragmatics frontier: disjunctions and free choice. *Cognition*, *130,* 380–396. doi:10.1016/j.cognition.2013.11.013

Chevallier, C., Bonnefond, M., Van der Henst, J.-B., & Noveck, I. (2010). Using ERPs to capture prosodic stress and inference making. *Italian Journal of Linguistics*, *22*(1), 125–152.

Chevallier, C., Noveck, I. A., Nazir, T., Bott, L., Lanzetti, V., & Sperber, D. (2008). Making disjunctions exclusive. *Quarterly Journal of Experimental Psychology*, *61*, 1741–1760. doi:10.1080/17470210701712960

Chevallier, C., Wilson, D., Happé, F., & Noveck, I. (2010). Scalar inferences in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *40*, 1104–1117. doi:10.1007/s10803-010-0960-8

Clark, H. H. & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, *3,* 472–517. doi:10.1016/0010-0285(72)90019-9

Cousineau, D. (2005). Confidence intervals in within-subject designs: a simpler solution to Loftus and Masson’s method. *Tutorials in Quantitative Methods for Psychology*, *1*, 42–45. doi:10.3758/BF03210951

Cremers, A., & Chemla, E. (2014). Direct and indirect scalar implicatures share the same processing signature. In S. Pistoia Reda (Ed.), *Pragmatics, semantics and the case of* *scalar implicatures* (pp. 201–227). London, United Kingdom: Palgrave Macmillan. doi:10.1057/9781137333285 8

De Neys, W., & Schaeken, W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology, 54*, 128–133. doi:10.1027/1618-3169.54.2.128

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: a constraint-based approach. *Cognitive Science, 39*, 667–710. doi:10.1111/cogs.12171

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: a visual world eye-tracking study. *Cognitive Science, 40*, 172–201. doi:10.1111/cogs.12227

Deschamps, I., Agmon, G., Loewenstein, Y., & Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition, 143,* 115–128. doi:10.1016/j.cognition.2015.06.006

Dieussaert, K., Verkerk, S., Gillard, E., & Schaeken, W. (2011). Some effort for some: Further evidence that scalar implicatures are effortful. *Quarterly Journal of Experimental Psychology, 64*, 2352–2367. doi:10.1080/17470218.2011.588799

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*, 1–38. doi:10.1163/187730909X12538045489854

Fauconnier, G. (1975). Polarity and the scale principle. *Proceedings of the Chicago Linguistics Society, 11*, 188– 199.

Feeney, A., Scrafton, S., Duckworth, A., & Handley, S. J. (2004). The story of *some*: everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology,* 58, 121–132. doi:10.1037/h0085792

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

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

Geurts, B., Katsos, N., Cummins, C., Moons, J., & Noordman, L. (2010). Scalar quantifiers: logic, acquisition, and processing. *Language and Cognitive Processes*, *25*, 130–148. doi:10.1080/01690960902955010

Glenberg, A. M., Robertson, D. A., Jansen, J. L., & Johnson-Glenberg, M. C. (1999). Not propositions. *Cognitive Systems Research*, *1*, 19–33. doi:10.1016/S1389-0417(99)00004- 2

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, volume 3: Speech acts* (pp. 41–58). New York, NY: Academic Press.

Gotzner, N. & Benz, A. (2018). The best response paradigm: a new approach to test implicatures of complex sentences. *Frontiers in Communication*, *2*, 21. doi:10.3389/fcomm.2017.00021

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). "Some," and possibly all, scalar inferences are not delayed: evidence for immediate pragmatic enrichment. *Cognition*, *116*, 42–55. doi:10.1016/j.cognition.2010.03.014

Heim, I. (2007). "Little". In M. Gibson & J. Howell (Eds.), *Proceedings of Semantics and Linguistic Theory 16* (pp. 35–58). Ithaca, NY: Cornell University. doi:10.3765/salt.v16i0.2941

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

Horn, L. R. (1972). *On the semantic properties of logical operators in English* (Unpublished doctoral dissertation). University of California, Los Angeles.

Horn, L. R. (1989). *A natural history of negation*. Chicago, IL: University of Chicago Press.

Huang, Y. T., & Snedeker, J. (2009). On-line interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, *58*, 376–415. doi:10.1016/j.cogpsych.2008.09.001

Huang, Y. T., & Snedeker, J. (2011). Logic and conversation revisited: Evidence for a division between semantic and pragmatic content in real-time language comprehension. *Language and Cognitive Processes*, *26*, 1161–1172. doi:10.1080/01690965.2010.508641

Huang, Y. T., & Snedeker, J. (2018). Some inferences still take time: prosody, predictability, and the speed of scalar implicatures. *Cognitive Psychology*, *102*, 105–126. doi:10.1016/j.cogpsych.2018.01.004

Just, M. A. & Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, *10*, 244-253. doi:10.1016/S0022-5371(71)80051-8

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

Kaup, B., Yaxley, R. H., Madden, C. J., Zwaan, R., & Lüdtke, J. (2007). Experiential simulations of negated text information. *Quarterly Journal of Experimental Psychology*, *60*, 976–990. doi:10.1080/17470210600823512

Kennedy, C. & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, *81*, 345-381. doi:10.1353/lan.2005.0071

Klatzky, R., Clark, E. V., & Macken, M. (1973). Asymmetries in the acquisition of polar adjectives: linguistic or conceptual? *Journal of Experimental Child Psychology*, *16*, 32–46. doi:10.1016/0022-0965(73)90060-X

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2013). lmerTest: tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package) [R package]. Retrieved from http://cran.r-project.org/package=lmerTest

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

Marty, P., & Chemla, E. (2013). Scalar implicatures: working memory and a comparison with only. *Frontiers in Psychology*, *4*, 1–12. doi:10.3389/fpsyg.2013.00403

Marty, P., Chemla, E., & Spector, B. (2013). Interpreting numerals and scalar items under memory load. *Lingua*, *133*, 152–163. doi:10.1016/j.lingua.2013.03.006

Matsumoto, Y. (1995). The conversational condition on Horn scales. *Linguistics and Philosophy*, *18*, 21–60. doi:10.1007/bf00984960

McLeod, A. (2011). Kendall: Kendall rank correlation and Mann-Kendall trend test [R package]. Retrieved from https://CRAN.R-project.org/package=Kendall

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General*, *130*, 621–640. doi:10.1037/0096-3445.130.4.621

Moxey, L. M., Sanford, A. J., & Dawydiak, E. J. (2001). Denial as controllers of negative quantifier focus. *Journal of Memory and Language*, *44*, 427–442. doi:10.1006/jmla.2000.2736

Moxey, L. M. (2006). Effects of what is expected on the focussing properties of quantifiers: a test of the presupposition-denial account. *Journal of Memory and Language*, *55*, 422–439. doi:10.1016/j.jml.2006.05.006

Nieuwland, M. S., Ditman, T., & Kuperberg, G. R. (2010). On the incrementality of pragmatic processing: an ERP investigation of informativeness and pragmatic abilities. *Journal of Memory and Language*, *63*, 324–346. doi:10.1016/j.jml.2010.06.005

Noveck, I. A. (2001). When children are more logical than adults: experimental investigations of scalar implicature. *Cognition*, *78*, 165–188. doi:10.1016/S0010-0277(00)00114-1

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: an evoked potentials study. *Brain and Language*, *85*, 203–210. doi:10.1016/S0093-934X(03)00053-1

Papafragou, A., & Musolino, J. (2003). Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition*, *78*, 253–282. doi:10.1016/S0010-0277(02)00179-8

Pew Research Center. (2016). Research in the crowdsourcing age, a case study [Report]. Retrieved from http://www.pewinternet.org/2016/07/11/research-in-the-crowdsourcing-age-a-case-study

Politzer-Ahles, S. & Fiorentino, R. (2013). The realization of scalar inferences: context sensitivity without processing cost. *PloS ONE, 8*, e63943. doi:10.1371/journal.pone.0063943

Politzer-Ahles, S., & Husband, M. E. (2018). Eye movement evidence for context-sensitive derivation of scalar inferences. *Collabra*, *1*, 1–13. doi:10.1525/collabra.100

Pouscoulous, N., Noveck, I. A., Politzer, G., & Bastide, A. (2007). A developmental investigation of processing costs in implicature production. *Language Acquisition*, *14*, 347–375. doi:10.1080/10489220701600457

R Development Core Team. (2006). R: a language and environment for statistical computing [Computer software manual]. Vienna, Austria.

Rees, A., & Bott, L. (2018). The role of alternative salience in the derivation of scalar implicatures. *Cognition*, *176*, 1–14. doi:10.1016/j.cognition.2018.02.024

Rips, L. J. (1975). Quantification and semantic memory. *Cognitive Psychology*, *7*, 307–340. doi:10.1016/0010-0285(75)90014-6

Romoli, J., & Schwarz, F. (2015). An experimental comparison between presuppositions and indirect scalar implicatures. In F. Schwarz (Ed.), *Experimental perspectives on presuppositions* (pp. 215–240). Cham, Germany: Springer. doi:10.1007/978-3-319-07980-6_10

Simons, M., & Warren, T. (2018). A closer look at the strengthened readings of scalars. *Quarterly Journal of Experimental Psychology*, *7*, 272–279. doi:10.1080/17470218.2017.1314516

Soames, S. (1982). How presuppositions are inherited: a solution to the projection problem. *Linguistic Inquiry*, *13*(3), 483–545.

Sperber, D., & Wilson, D. (1986). *Relevance: communication and cognition*. Oxford, United Kingdom: Blackwell.

Sperber, D., & Wilson, D. (1987). Précis of relevance: communication and cognition. *Behavioral and Brain Sciences*, *10*, 697–754. doi:10.1017/S0140525X00055345

Tian, Y., Breheny, R., & Ferguson, H. (2010). Why we simulate negated information: a dynamic pragmatic account. *Quarterly Journal of Experimental Psychology*, *63*, 2305–2312. doi:10.1080/17470218.2010.525712

Tieu, L., Yatsushiro, K., Cremers, A., Romoli, J., Sauerland, U., & Chemla, E. (2017). On the role of alternatives in the acquisition of simple and complex disjunctions in French and Japanese. *Journal of Semantics, 34*, 127–152. doi:10.1093/jos/ffw010

Tomlinson Jr., J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: scalar implicatures are understood in two steps. *Journal of Memory and Language, 69*, 18–35. doi:10.1016/j.jml.2013.02.003

van Tiel, B., Kissine, M., & Noveck, I. (2018). Reasoning with 'some'. *Journal of Semantics*. Advance online publication. doi:10.1093/jos/ffy012

van Tiel, B., & Schaeken, W. (2016). Processing conversational implicatures: Alternatives and counterfactual reasoning. *Cognitive Science, 41*, 1–36. doi:10.1111/cogs.12362

van Tiel, B., van Miltenburg, E., Zevakhina, N., & Geurts, B. (2016). Scalar diversity. *Journal of Semantics, 33*, 137–175. doi:10.1093/jos/ffu017

Ward Jr., J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*, 236–244. doi:10.1080/01621459.1963.10500845

Wason, P. (1965). The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior*, *4*, 4–11. doi:10.1016/s0022-5371(65)80060-3