



Adjectival polarity and the processing of scalar inferences

BOB VAN TIEL 

ELIZABETH PANKRATZ 

**Author affiliations can be found in the back matter of this article*

RESEARCH

]u[ubiquity press

Abstract

In a seminal study, Bott & Noveck (2004) found that the computation of the scalar inference of 'some' implying 'not all' was associated with increased sentence verification times, suggesting a processing cost. Recently, van Tiel and colleagues (2019b) hypothesised that the presence of this processing cost critically depends on the polarity of the scalar word. We comprehensively evaluated this polarity hypothesis on the basis of a sentence-picture verification task in which we tested the processing of 16 types of adjectival scalar inferences. We develop a quantitative measure of adjectival polarity which combines insights from linguistics and psychology. In line with the polarity hypothesis, our measure of polarity reliably predicted the presence or absence of a processing cost (i.e., an increase in sentence verification times). We conclude that the alleged processing cost for scalar inferencing in verification tasks is not due to the process of drawing a scalar inference, but rather to the cognitive difficulty of verifying negative information.

CORRESPONDING AUTHOR:

Bob van Tiel

Donders Institute for Brain,
Cognition and Behaviour,
Postbus 9010, 6500 GL
Nijmegen, NL

bobvantiel@gmail.com

KEYWORDS:

scalar inference; adjective;
polarity; sentence processing;
implicature

TO CITE THIS ARTICLE:

van Tiel, Bob and Elizabeth
Pankratz. 2021. Adjectival
polarity and the processing
of scalar inferences. *Glossa: a
journal of general linguistics*
6(1): 32. 1–21. DOI: [https://
doi.org/10.5334/gjgl.1457](https://doi.org/10.5334/gjgl.1457)

An utterance of (1a) can be interpreted in (at least) two ways.

- (1) a. It is warm outside.
b. It is hot outside.

On its *one-sided* interpretation, the utterance conveys that the temperature outside exceeds some contextually determined value, e.g., 20 degrees Celsius. On its *two-sided* interpretation, the utterance conveys, in addition, that the temperature lies below another contextually determined value, e.g., 30 degrees Celsius. In other words, on its two-sided interpretation, an utterance of (1a) conveys that (1b) is false.

Most current theories assume that the one-sided interpretation corresponds to the *literal* interpretation of (1a). To explain how the two-sided interpretation emerges from this literal interpretation, it is generally assumed that words like ‘warm’ evoke lexical scales consisting of words that are ordered in terms of logical strength, e.g., ⟨warm, hot⟩. Here, ‘hot’ is assumed to be logically stronger than ‘warm’ (at least at the level of literal meaning) since it refers to a more restrictive range of situations. For example, 20 degrees Celsius counts as warm but not hot, but there are no situations that count as hot but not warm (again: at the level of literal meaning). Given a lexical scale, uttering a sentence containing the weaker scalar word may imply that the corresponding sentence containing the stronger scalar word is false. Hence, these inferences have become known as *scalar inferences* (e.g., Horn 1972; Gazdar 1979; Soames 1982; Geurts 2010; Huang 2014).

Scalar inferences are commonly explained as a variety of *conversational implicature*, i.e., as a type of inference that can be calculated on the basis of the literal interpretation of an utterance and the assumption that the speaker is cooperative (Grice 1975). In the case at hand, someone who utters (1a) could have been more informative—and therefore cooperative—by saying (1b). Why didn’t she? Presumably because she believes that it is not hot outside, i.e., she believes that (1b) is false.

According to this implicature-based explanation, the one-sided interpretation is theoretically prior to the two-sided interpretation, since the one-sided interpretation serves as a premise in the reasoning process that ultimately leads to the scalar inference (and, consequently, the two-sided interpretation). An important question is whether the theoretical priority of the literal interpretation is reflected in listeners’ cognitive processing, i.e., whether the computation of scalar inferences is associated with a *processing cost* vis-à-vis the literal interpretation, in line with the latter’s theoretical priority (Récanati 1995).

Levinson (2000) explicitly rejects such an isomorphism between derivational complexity and processing difficulty. Levinson’s point of departure is the observation that human communication has a comparatively slow information transmission rate because of the time needed for phonetic articulation (i.e., we can only talk so fast). One way of reducing this articulatory bottleneck is by incorporating certain pragmatic inferences—including scalar inferences—into the lexical meaning. Levinson argues that this process of lexical integration is pragmatic, but occurs automatically during the construction of the initial interpretation of the utterance. Thus, according to Levinson, an utterance of (1a) receives a scalar inference by default, though this inference can be overridden in certain special situations (e.g., when the speaker continues with ‘In fact, it is hot outside’).

Proponents of *relevance theory* take a more nuanced stance on the cognitive cost of scalar inferencing (e.g., Sperber & Wilson 1987; 1995; Noveck & Sperber 2007; Chevallier et al. 2008). According to relevance theory, listeners try to piece together the speaker’s intention based on the literal interpretation of an utterance, the surrounding context, and the expectation that the utterance is optimally relevant. Relevance theorists argue that, if the context makes the two-sided interpretation sufficiently relevant (e.g., when (1a) is said to someone who wants to know what to wear today), scalar inferences may be computed without any processing cost, and the two-sided interpretation may even be easier to retrieve than the literal interpretation. However, if there is no such facilitating context—as will generally be the case in the experiments that we describe below—the literal interpretation of an utterance is a good first guess as to the

speaker's intention, and deriving the scalar inference involves an inferential process of meaning construction that is cognitively taxing and time-consuming.

Several more recent proposals side with relevance theory in assuming that the presence of a processing cost for scalar inferencing varies with certain methodological and contextual factors. However, they do not necessarily commit to the relevance-theoretic assumption that relevance is paramount in deciding whether or not a processing cost will be observed. Thus, e.g., these proposals have argued that the presence of a processing cost depends on the question under discussion (Westera 2017; Ronai & Xiang 2020), the structural characteristics of the alternatives (Chemla & Bott 2014; van Tiel & Schaeken 2016), the naturalness of the utterance (Degen & Tanenhaus 2016), and, as we will discuss in much more detail later, the *polarity* of the scalar inference (van Tiel et al. 2019b).

Testing these different theories about the processing of scalar inferences requires operationalising the notion of a processing cost. Various proposals have been made in this respect, focusing on participants' eye movements (e.g., Grodner et al. 2010; Huang & Snedeker 2018), brain signals (e.g., Noveck & Posada 2003; Barbet & Thierry 2018), reading times (e.g., Breheny et al. 2006; Politzer-Ahles & Husband 2018), and working memory capacity (e.g., De Neys & Schaeken 2007; Marty & Chemla 2013). In this study, we focus on the idea that processing costs can be measured by looking at sentence verification times. In the next section, we briefly discuss previous studies using this measure. We show that these studies have given rise to conflicting results, and describe a recent proposal that aims to make sense of these conflicting data in terms of the polarity of scalar words. Afterwards, we turn to our own study in which we systematically and extensively tested the polarity-based explanation.

1.1 Previous sentence verification studies

In sentence verification studies, participants are presented with a sentence and have to decide whether that sentence is true or false in a given situation. This situation can be presented pictorially or correspond to participants' world knowledge. To carry out the verification process, it is often assumed that participants represent both the sentence and the situation in a common format, e.g., a proposition. In addition, participants initialise a truth index that tracks the truth value of the sentence. Sentence verification then consists in systematically manipulating and comparing the representations associated with the sentence and the situation, and carrying out operations on the truth index (cf. Clark & Chase 1972; Carpenter & Just 1975).

To examine whether the computation of scalar inferences is associated with a processing cost, Bott & Noveck (2004) tested the ⟨some, all⟩ scale in a series of sentence verification tasks. Participants in their experiments had to indicate the truth value of underinformative sentences like (2).

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

These sentences are true when interpreted literally, since, e.g., there are dogs that are mammals, but they are false when the scalar inference is computed and 'some' is interpreted as 'some but not all', since, in fact, all dogs are mammals. Hence, participants' truth judgements to these underinformative sentences are indicative of whether or not they computed a scalar inference.

In Bott and Noveck's Exp. 3, participants gave intuitive truth judgements to sentences such as (2). Many participants were ambivalent about the truth of underinformative sentences like these, varying their responses across structurally similar trials. Comparing the verification times of these ambivalent participants, Bott and Noveck found that it took participants significantly longer to answer 'false' (i.e., the answer suggesting a two-sided interpretation) than 'true' (i.e., the answer suggesting a literal interpretation). This difference in verification times was absent in a control condition with sentences that were unambiguously true or false, as in (3).

- (3) a. Some mammals are dogs.
 b. Some dogs are birds.

The pattern of results that Bott and Noveck observed suggests that the computation of scalar inferences is associated with a processing cost, at least in out-of-the-blue contexts. This

conclusion is in line with relevance theory and several more recent approaches (e.g., Chemla & Bott 2014; van Tiel & Schaeken 2016; Degen & Tanenhaus 2019), but speaks against Levinson's proposal that the default interpretation of 'some' is two-sided.

In what follows, we refer to Bott and Noveck's finding that participants take significantly longer to reject underinformative sentences like (2) than to accept them as the *B&N effect*. The B&N effect for the ⟨some, all⟩ scale has been replicated in numerous studies (e.g., Noveck & Posada 2003; Tomlinson Jr. et al. 2013; Chemla & Bott 2014; Cremers & Chemla 2014; van Tiel & Schaeken 2016; Ronai & Xiang 2020). At the same time, however, several studies have shown that the B&N effect does not always generalise beyond the specific case of 'some'.

For example, Chevallier and colleagues (2010) tested the ⟨or, and⟩ scale in a sentence-picture verification task. Participants in their experiment had to judge the truth value of sentences like (4) in displays showing different types of objects.

(4) There is a sun or a train.

In the target condition, the display for (4) showed both a sun and a train. Here, the sentence is literally true but false if the scalar inference is computed and 'or' is interpreted as excluding 'and'. As in Bott and Noveck's study, many participants vacillated between responding with 'true' or 'false' in the target condition. Unlike Bott and Noveck's study, however, Chevallier and colleagues did not observe a significant difference in verification times between 'true' and 'false' answers.

Even more challenging data comes from studies testing the processing of scalar words in negative sentences, as in (5) (Cremers & Chemla 2014; Romoli & Schwarz 2015; Marty et al. 2020).

(5) a. Not all dogs are insects.
b. Not all parrots are mammals.

On their literal interpretation, the sentences in (5) merely convey that there are dogs that are not insects, and that there are parrots that are not mammals. So on their literal interpretation, these sentences are true. However, the sentences in (5) may give rise to the *indirect* scalar inference that the corresponding sentences with 'some' are false, i.e., that at least some dogs are insects, and that at least some parrots are mammals. Clearly, these scalar inferences are false.

Cremers & Chemla (2014: Exp. 1) asked participants to give their intuitive truth judgements to sentences such as (5). As in Bott and Noveck's study, participants were ambivalent about the truth of these underinformative sentences, responding differently across structurally similar trials. However, unlike Bott and Noveck's study, Cremers and Chemla found that participants were *faster* when responding 'false' than when responding 'true' (recall that the B&N effect consists in *slower* verification times when responding 'false'). This difference in response times was absent in a control condition involving sentences that were unambiguously true or false, as in (6).

(6) a. Not all mammals are dogs.
b. Not all dogs are mammals.

Romoli & Schwarz (2015) and Marty et al. (2020) found the same pattern of response times for various other types of indirect scalar inferences. These findings are noteworthy because they suggest that scalar inferences are processed differently when negation is involved. We will return to this point below.

To obtain a more comprehensive picture of the generalisability of the B&N effect, van Tiel and colleagues (2019b) tested the processing of seven lexical scales: ⟨some, all⟩, ⟨or, and⟩, ⟨most, all⟩, ⟨might, must⟩, ⟨try, succeed⟩, ⟨low, empty⟩, and ⟨scarce, absent⟩. In their Exp. 1, participants gave intuitive truth judgements to sentences containing the weaker scalar word. These sentences were presented in two types of displays. In control displays, the sentence was unambiguously true or false. In target displays, the sentence was literally true but false if the corresponding scalar inference was computed. To illustrate, [Table 1](#) shows the materials for ⟨some, all⟩ and ⟨low, empty⟩.




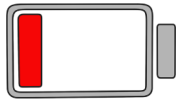
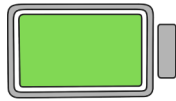

Sentence	Control-True	Control-False	Target
Some of the socks are pink.			
The battery is low.			

Table 1 Materials used by van Tiel et al. (2019b: Exp. 1) for the lexical scales (some, all) and (low, empty).

In line with Bott and Noveck’s study, van Tiel and colleagues found that, in the case of ‘some’, participants were significantly slower to answer ‘false’ than ‘true’ in the target condition, whereas no difference in verification times was observed in the control condition. Van Tiel and colleagues also observed a B&N effect for ‘or’ (in contrast with the aforementioned study by Chevallier et al. 2010), ‘might’, ‘most’, and ‘try’. In the case of ‘low’ and ‘scarce’, however, no significant difference in verification times between ‘true’ and ‘false’ responses was observed.

To explain this pattern of results, van Tiel and colleagues rely on the notion of *polarity*. In particular, van Tiel and colleagues argue that only the scalar inferences associated with *positive* scalar words are associated with a B&N effect, and that this effect is the result of the cognitive difficulty of processing the corresponding *negative* scalar inference. In the next section, we first introduce the notion of polarity. Afterwards, we discuss in more detail the polarity-based explanation proposed by van Tiel and colleagues.

1.2 Polarity

Polarity is a fundamental but multifarious construct that refers to the fact that some words in natural language are positive while others are negative (cf. Horn 1989: Ch. 1–3 for an excellent overview). For example, ‘warm’ is usually assumed to be positive, whereas ‘cold’ is assumed to be negative. As this example already shows, negative words are not always explicitly marked for negativity. When negative marking is absent, these words are assumed to have an implicit negative element in their underlying semantic representation (Clark 1974; Heim 2008; Moracchini 2019). The notion of polarity has been prominently studied in linguistics and psychology; mostly disparately, but cf. Ingram et al. (2016) and Nouwen (2020) for more integrative approaches. However, these two fields have operationalised polarity in importantly different ways.

In linguistics, polarity is usually operationalised in terms of *markedness*, i.e., negative words tend to be marked compared to their positive counterparts (e.g., Greenberg 1966; Lyons 1968; Clark & Clark 1977; Givón 1979; Lehrer & Lehrer 1982; Lehrer 1985; Sassoon 2010; Morzycki 2015). There are various ways of determining whether or not a word is marked. One such way relies on the fact that certain words make reference to a measurement scale in their semantics (e.g., Kennedy & McNally 2005; Solt 2015). To illustrate, compare ‘many’ and ‘few’. Both of these words operate on the quantity scale. However, whereas ‘many’ denotes a *lower* bound on the quantity scale (e.g., ‘Many flowers are red’ implies that the number of red flower is *greater* than a contextually determined threshold), ‘few’ denotes an *upper* bound (van Tiel et al. 2021). To put it differently, “many-ness” and quantity are positively related, whereas “few-ness” and quantity are negatively related. As a consequence, ‘many’ is usually characterised as positive and ‘few’ as negative.

Van Tiel and colleagues rely on this characterisation, which they call the *scalarmity criterion*, to intuitively classify the scalar words in their sample as positive or negative. Based on this criterion, they labelled ‘low’ and ‘scarce’ as negative, and all other words as positive. Recall that ‘low’ and ‘scarce’ were also the only two scalar words that failed to give rise to the B&N effect. This concurrence between polarity and processing led van Tiel and colleagues to hypothesise that polarity is the key feature in determining whether or not a B&N effect will be observed.

However, in addition to the scalarity criterion, there are various other diagnostics of linguistic polarity. Since adjectives are of particular interest for the current study, we focus here on two ways of diagnosing the linguistic polarity of adjectives. Both of these diagnostics build on a standard assumption in linguistics that many adjectives are members of antonym pairs where one member is positive and the other is negative (e.g., Lehrer & Lehrer 1982).

A first way of diagnosing adjectival polarity involves the interpretation of ‘how’ questions. In particular, ‘how’ questions involving positive adjectives tend to be neutral, whereas those involving negative adjectives tend to presuppose that the adjective holds (e.g., Rett 2008). To illustrate, consider the questions in (7).

- (7) a. How long is a day on Venus?
b. How short is a day on Venus?

Whereas (7a) is neutral about whether days on Venus are long or short, (7b) intuitively suggests that the speaker believes they are short. This observation suggests that ‘long’ is positive, while ‘short’ is negative. A direct consequence of the fact that negative adjectives are biasing in ‘how’ questions is that they are less likely to occur in such questions than positive adjectives—e.g., the phrase ‘how long’ is much more frequent in the ENCOW16A corpus (a web corpus consisting of almost 17 billion tokens, cf. Schäfer & Bildhauer 2012; Schäfer 2015) than the phrase ‘how short’ (199,033 vs. 2,456 occurrences).

A second way of linguistically delineating positive and negative adjectives looks at ratio phrases, such as ‘twice as’ and ‘half as’ (cf. Sassoon 2010). Ratio phrases presuppose a natural zero point. For many positive adjectives, such as ‘tall’ and ‘old’, such a natural zero point is intuitively available. For example, conceptually, there is such a thing as zero tallness (i.e., being 0 centimeters tall) or zero oldness (i.e., being 0 days old). For many negative adjectives, however, there is no natural zero point. Thus, there is no such thing as zero shortness (which would correspond to infinite length) or zero youngness (infinite age). As a consequence, the positive adjective ‘old’ is felicitous in ratio phrases, while its negative counterpart ‘young’ is slightly odd, as shown by the minimal pair in (8).

- (8) a. She is twice as old as him.
b. ?He is twice as young as her.

In line with this observation, Sassoon (2010) provides corpus data showing that positive adjectives are—as a rule—significantly more frequent than negative adjectives in ratio constructions such as ‘twice as’. In line with these data, ‘twice as old’ was substantially more frequent than ‘twice as young’ in the ENCOW16A corpus (258 vs. 3 occurrences).

In psychology, polarity is usually defined in terms of *subjective valence*, i.e., in terms of the positive or negative connotations that people have with a particular word (e.g., Wason 1959; Osgood & Richards 1973). To illustrate, Mohammad (2018) presented participants with short lists of words and asked them to rank these based on their valence. These rankings were then converted to numeric values between 0 (indicating that the word was always ranked at the bottom of the list) and 1 (always ranked at the top). Thus, e.g., ‘good’ was associated with a value of 0.938; ‘bad’ with a value of 0.125, which reflects the intuition that ‘good’ is positive and ‘bad’ is negative.

The psychological notion of polarity as subjective valence also reverberates in natural language in several ways (e.g., Boucher & Osgood 1969; Benjafield & Adams-Webber 1976; Paradis et al. 2012). For example, it has been found that psychologically positive words are more frequently attested than negative ones (i.e., the *Polyanna hypothesis* formulated and tested by Boucher & Osgood 1969). In line with this idea, a search in the ENCOW16A corpus shows that ‘good’ is more than four times as frequent as ‘bad’ (10,869,258 vs. 2,289,838 occurrences).

In most cases, the linguistic and psychological notions of polarity go hand in hand, but not always. For example, as we just saw, from a linguistic perspective, ‘old’ is positive while ‘young’ is negative. From a psychological perspective, the converse holds: ‘young’ is positive and ‘old’ is negative (e.g., in the study by Mohammad 2018, participants gave ‘old’ a valence rating of 0.41 and ‘young’ a rating of 0.81). Therefore one of the main contributions of this paper is the synthesis of several (potentially conflicting) polarity diagnostics into a single continuous polarity measure. We use this polarity measure to systematically test the speculation from van Tiel and colleagues that only positive scalar words are associated with a B&N effect (recall that they relied on the scalarity criterion, a rather intuitive measure, to classify scalar words as positive or negative). However, before going into more detail about the present study, we explore why it is plausible that polarity affects the processing of scalar inferences.

1.3 A polarity-based explanation

In order to explain why only positive scalar words are associated with a B&N effect, van Tiel and colleagues rely on the observation that verification times are systematically affected by the polarity of the sentence. To illustrate, consider the three sentences in (9). These sentences vary in their polarity: (9a) is positive, (9b) contains the implicitly negative word ‘below’, and (9c) contains the explicit sentential negation ‘not’. In what follows, we will conveniently refer to these three types of sentences as *positives*, *implicit negatives*, and *explicit negatives*.

- (9) a. The star is above the cross.
b. The cross is below the star.
c. The cross is not above the star.

Clark & Chase (1972) measured the verification times for these three types of sentences in displays that always showed two vertically juxtaposed images. Crucially, the three sentences in (9) are all equivalent in these displays. Nonetheless, Clark and Chase found that participants were significantly faster to verify positives like (9a) than implicit negatives like (9b), which, in turn, were verified significantly more rapidly than explicit negatives like (9c). In other words, Clark and Chase found evidence for the generalisation in (10), where ‘<’ denotes *faster* verification times.

- (10) positive < implicit negative < explicit negative

Clark and Chase’s findings have been replicated in numerous studies (e.g., Wason 1972; Carpenter & Just 1975; Fodor et al. 1975; Cheng & Huang 1980; Proctor & Cho 2006). However, it should be noted that all of these studies tested sentences that were presented without any relevant context. Indeed, later work has convincingly shown that these findings do not always generalise to more contextualised settings. For example, several studies found that explicit negatives can be verified as fast as positives if the context is right (cf. Wason 1965; Nieuwland & Kuperberg 2008; Tian et al. 2010). In what follows, we will largely ignore this important qualification, since almost all verification studies on the processing of scalar inferences have made use of out-of-the-blue contexts (but cf. Ronai & Xiang 2020 for a recent exception).

To see how the generalisation in (10) may explain van Tiel and colleagues’ observation that only positive scalar words are associated with a B&N effect, consider their target sentence for the positive scalar word ‘some’ and its scalar inference in (11).

- (11) Some of the socks are pink.
↪ Not all of the socks are pink.

Participants who interpreted the target sentence literally only had to verify a positive sentence, whereas participants who arrived at a two-sided interpretation also had to verify the explicitly negative scalar inference. Given the generalisation in (10), we may expect that verifying the explicitly negative scalar inference leads to elevated response times compared to the literal interpretation. So, even if we assume that participants who arrived at a two-sided interpretation of the target sentence verified the literal interpretation and the scalar inference in parallel, it follows that they are expected to be significantly slowed down compared to participants who interpreted the sentence literally.

Now consider the target sentence for the implicitly negative scalar word ‘scarce’ along with its scalar inference in (12).

- (12) Red flowers are scarce.
↪ Red flowers are not absent.

Participants who arrived at a literal interpretation of the target sentence had to verify an implicitly negative sentence, since ‘scarce’ is implicitly negative. What about the scalar inference? Superficially, the scalar inference appears to involve a double negation in ‘not absent’. Hence, intuitively, one might suppose that the verification of the scalar inference should take longer than verifying the literal interpretation. However, van Tiel and colleagues contend that the scalar inference in this case is verified at least as fast as the literal interpretation. There are various arguments that may support this proposal.

First, it has been found that, in at least some cases, sentences containing two negative elements are processed more rapidly than the corresponding sentences with a single negative element. For example, Sherman (1976) found that participants were faster to verify sentences containing the double negation ‘no one doubted’ (‘doubt’ is implicitly negative) than sentences containing just the single negative word ‘doubted’ in an otherwise positive sentence. Hence, it could be that ‘not absent’ is easier to verify than ‘scarce’. Second, it could be the case that the scalar inference is cognitively represented as a positive, i.e., as ‘Red flowers are present’. The double negation could be eliminated on the fly, or, perhaps more plausibly from a psychological standpoint, it could be that the positive form of the scalar inference is directly associated with the scalar word, similarly to Levinson’s (2000) defaultist proposal, so that participants who derive the scalar inference interpret the target sentence in (12) as ‘Red flowers are scarce but present’.

In any case, if we assume that the scalar inference in (12) can be verified at least as rapidly as the literal interpretation, and if we furthermore assume that participants who compute the scalar inference verify both the literal interpretation and the scalar inference in parallel, it follows that participants who computed the scalar inference should be equally fast as participants who arrived at a literal interpretation. One might even expect participants who derived the scalar inference to be faster than participants who arrived at the literal interpretation, since the positivity of the scalar inference seems to entail that it should be verified faster than the implicitly negative literal interpretation, and since the sentence may be judged false as soon as the scalar inference is verified. However, psycholinguistic studies have consistently shown that ‘false’ responses to positive sentences are generally slightly delayed compared to ‘true’ responses, which might mitigate that verification time advantage for positives relative to implicit negatives, and thus lead to roughly equal verification times for literal and two-sided interpretations (e.g., Clark & Chase 1972).

This polarity-based explanation also harmonises with the previously discussed findings on indirect scalar inferences. To illustrate, (13) shows a target sentence from Cremers and Chemla’s (2014) study, as well as its scalar inference.

- (13) Not all dogs are insects.
 ↗ It’s not the case that not some dogs are insects.

Again, the proposal is that the scalar inference is verified more rapidly than the literal interpretation, either because the double negation is eliminated or because ‘not all’ is statistically associated with ‘some’ (rather than the equivalent ‘not not some’). The reason that (13) gave rise to the reverse B&N effect—rather than the absence of any effect, as van Tiel and colleagues found for cases like ‘scarce’—is that the target sentence contains an explicit negation. As noted before, explicit negatives take longer to verify than implicit negatives. Hence, participants who accepted the target sentence had to verify the more time-consuming literal interpretation, whereas participants who arrived at the two-sided interpretation verified both the literal interpretation and the (positive) scalar inference in parallel. In the latter case, participants could respond with ‘false’ as soon as they realised that the scalar inference was false, which took less time than verifying the literal interpretation and responding with ‘true’.

Table 2 succinctly summarises the predictions of the polarity-based explanation. If this explanation is on the right track, it would mean that the B&N effect does not reflect a processing cost for scalar inferencing, since the scalar inferences of negative scalar words are not associated with a B&N effect. Indeed, if correct, it would mean that the B&N effect is only reflective of more general processing difficulties associated with the verification of negative information relative to positive information.

<i>Scalar word</i>	<i>Literal interpretation</i>	<i>Scalar inference</i>	<i>B&N effect</i>
positive (e.g., ‘some’)	positive	expl. negative	present
impl. negative (e.g., ‘scarce’)	impl. negative	positive	absent
expl. negative (e.g., ‘not all’)	expl. negative	positive	reversed

Table 2 Predictions of the polarity-based explanation about the polarity properties of the literal interpretation and the scalar inference, and about the B&N effect.

However, the current support for the polarity hypothesis is comparatively thin, consisting solely of the data for ‘low’ and ‘scarce’ (as well as perhaps earlier data on indirect scalar inferences). Moreover, in addition to being the only negative scalar words tested by van Tiel and colleagues, ‘low’ and ‘scarce’ were also the only *adjectival* scalar words they tested. This may have influenced the results in various ways, e.g., one can imagine that adjectival scales are less salient given the openness of the grammatical class (cf. van Tiel et al. 2016).

In this study, we test the hypothesis that only positive scalar words give rise to the B&N effect in a more comprehensive and systematic way by investigating the processing of 16 adjectival scalar words of both positive and negative polarity. Rather than relying on one subjective diagnostic to classify scalar words in a binary way as either positive or negative, we combined the outcomes of four objectively measurable diagnostics for polarity to obtain a gradient measure of polarity. Consequently, we tested whether this gradient measure of polarity predicted the presence or absence of a B&N effect. In the next section, we describe our study in more detail.

1.4 Our study

Our study tested 16 adjectival scales: ⟨ajar, open⟩, ⟨breezy, windy⟩, ⟨chubby, fat⟩, ⟨content, happy⟩, ⟨cool, cold⟩, ⟨drizzly, rainy⟩, ⟨fair, good⟩, ⟨low, empty⟩, ⟨mediocre, bad⟩, ⟨passable, good⟩, ⟨ripe, overripe⟩, ⟨scarce, absent⟩, ⟨sleepy, asleep⟩, ⟨unlikely, impossible⟩, ⟨warm, hot⟩, and ⟨youthful, young⟩. For each scale, we constructed a simple sentence containing the weaker scalar word, and, for each sentence, we created three images: a target image where the sentence was literally true but where its scalar inference was false, and two control images where the sentence was unambiguously true or false. See the Appendix for an overview of the sentences and images that we tested.

Participants in the experiment first saw the sentence. Once they finished reading the sentence, they pressed the space bar to see the image. At that point, they had to indicate whether they felt the sentence they had just read was a good or bad description of the corresponding image. We measured their verification times (i.e., the time between image onset and the point at which one of the response buttons was pressed) to establish the presence or absence of a B&N effect, i.e., to determine whether or not verification times were slower for ‘false’ than for ‘true’ answers in the target condition, vis-à-vis the control condition.

To test the polarity hypothesis—i.e., the idea that only positive scalar words give rise to the B&N effect—we had to determine the polarity of the scalar words in our study. Here, we focus on the stronger word on the scale, since the polarity-based explanation crucially makes reference to the polarity of the negated alternative, though if the literature is right, all words on a scale should share the same polarity (i.e., the *scalarity* constraint, cf. Fauconnier 1975; Horn 1989). In the previous section, we discussed five diagnostics that can be used to determine the polarity of adjectives:

- i. *Scalarity*: Positive words denote a lower bound on their measurement scale.
- ii. *Questions*: Positive adjectives are neutral in ‘how’ questions.
- iii. *Ratio*: Positive adjectives are more felicitous in ratio phrases like ‘twice as’.
- iv. *Valence*: Positive words are judged as having more positive connotations.
- v. *Frequency*: Positive words are more frequent than negative ones.

The first three diagnostics reflect the linguistic notion of polarity as markedness; the last two diagnostics reflect the psychological notion of polarity as subjective valence.

Van Tiel and colleagues focused on the scalarity criterion in their study. However, in our study, we wanted to avoid using this criterion for two reasons. First, not all of the scalar words that we tested make reference to a clearly identifiable measurement scale. For example, in the case of ‘open’, it is unclear whether the underlying measurement scale is about openness or closedness. Second, and relatedly, the scalarity criterion crucially relies on researchers’ intuitions, which are not always reliable.

Rather than relying on one specific construal of polarity, or even one specific diagnostic measure, we made use of each of the remaining four diagnostics in the list. Unlike the scalarity criterion, these four diagnostics can be operationalised using objective data. We assume here that each

of the four diagnostics offers an approximation of a fundamental latent construct of polarity, and that, by combining these diagnostics, we are able to obtain a relatively reliable estimate of that construct. Crucially, we assume that polarity is gradient rather than binary; that is, words can be positive or negative to varying degrees. In making this decision, we do not want to question the value of focusing on one specific construal of polarity, e.g., in terms of markedness. However, even in the extensive line of work focusing on linguistic polarity, it has been observed that there is no fail-proof way of establishing the polarity of a word that consistently accords with linguists’ intuitions (e.g., Rett 2008; Sassoon 2010; Gotzner et al. 2018). By combining data from different diagnostics, we may mitigate potentially counterintuitive outcomes from any single diagnostic.

Hence, for each of the stronger scalar words in our experiment—as well as their corresponding antonyms—we obtained four measures, corresponding to the last four diagnostics in the list above: (i) their frequency in the phrase ‘how [adjective]’, (ii) their frequency in the phrases ‘twice as [adjective]’ and ‘half as [adjective]’, (iii) their valence ratings as reported by Mohammad (2018), and (iv) their overall frequency. The corpus counts for (i), (ii), and (iv) were taken from the ENCOW16A corpus (Schäfer & Bildhauer 2012; Schäfer 2015), and the counts for (i) and (ii) were relativised to the adjectives’ overall frequency. The corpus frequencies were always logarithmised as a way of reducing skewness. The outcome of each measure for the stronger scalar word was divided by the outcome of that measure for its antonym. Thus, values greater than 1 indicate that the stronger scalar word was positive relative to its antonym; values between 0 and 1 indicate the reverse. These resulting ratio values are provided in [Table 3](#).

Scale	Antonym	Question	Ratio	Valence	Frequency	Polarity
(ajar, open)	closed	1.52	1.00	2.58	1.09	1.18
(breezy, windy)	calm	1.00	0.38	0.86	0.74	-0.99
(chubby, fat)	skinny	1.00	0.59	1.21	1.28	1.06
(content, happy)	sad	1.10	2.11	4.44	1.11	2.18
(cool, cold)	hot	0.99	0.79	0.81	0.99	-0.36
(drizzly, rainy)	dry	0.35	1.00	1.68	0.81	-1.27
(fair, good)	bad	1.06	1.13	7.50	1.11	2.12
(low, empty)	full	0.89	1.00	0.31	0.87	-1.57
(mediocre, bad)	good	0.94	0.88	0.13	0.90	-0.69
(passable, good)	bad	1.06	1.13	7.50	1.11	2.12
(ripe, overripe)	unripe	1.00	1.00	0.72*	0.96	-0.28
(scarce, absent)	present	0.69	1.00	0.24	0.80	-1.31
(sleepy, asleep)	awake	0.82	1.00	0.91	1.04	-0.03
(unlikely, impossible)	possible	1.12	0.00	0.22	0.88	-1.83
(warm, hot)	cold	1.00	1.27	1.23	1.01	0.46
(youthful, young)	old	0.85	0.20	1.98	0.97	-0.85

Table 3 Lexical scales tested in the experiment and antonyms of the stronger scalemate. *Question*: Relative frequency of ‘how [adjective]’ in the ENCOW16A corpus (Schäfer & Bildhauer 2012; Schäfer 2015). *Ratio*: Relative frequency of ‘twice as [adjective]’ and ‘half as [adjective]’ in the ENCOW16A corpus. *Valence*: Relative subjective valence rating (Mohammad 2018). *Frequency*: Relative overall frequency in the ENCOW16A corpus. *Polarity*: Polarity value based on the first principal component of a principal component analysis on the basis of the values in Question, Ratio, Valence, and Frequency. Missing values due to zero or singular counts were set to 1 and are italicised in the table. *Neither ‘overripe’ nor ‘unripe’ was tested by Mohammad (2018); we used the valence ratings for the words ‘rotten’ and ‘raw’ instead.

Next, we carried out a principal component analysis based on these ratio values.¹ Principal component analyses are commonly used when trying to extract values from a latent parameter (in the case at hand: polarity) based on values from observable parameters that are assumed to approximate the latent parameter (in the case at hand: the values from the four diagnostics). A principal component analysis allows us to reduce the values from the observable parameters to a single value (i.e., the first principal component) in such a way that as much of the variance in the observable parameters is accounted for as possible. In our case, the first principal component explained 47% of the variance.

¹ This excellent idea was suggested to us by one of our anonymous reviewers.

The values on the first principal component are shown in [Table 3](#). Positive values stand for positive polarity; negative values for negative polarity. Encouragingly, values from the first principal component generally accord with our intuitions. ‘Happy’, ‘good’, and ‘open’ were assigned positive polarity values, since they received positive values on all four diagnostics; ‘rainy’, ‘windy’, and ‘absent’ received negative polarity values. Less clear-cut cases such as ‘override’, ‘asleep’, and ‘hot’ were somewhere in between.

We used the values from the first principal component to test the polarity hypothesis, which predicts that the B&N effect should interact with polarity. Recall that the B&N effect consists of slower verification times when responding ‘false’ than ‘true’ in the target condition vis-à-vis the control condition; i.e., the B&N effect consists in an interaction effect on verification times between condition (target vs. control) and response (‘true’ vs. ‘false’). Hence, the polarity hypothesis predicts a significant three-way interaction between condition (target vs. control), response (‘true’ vs. ‘false’), and polarity, such that the relative increase in verification times for ‘false’ compared to ‘true’ responses in the target condition increases with increasing polarity of the adjective.

The next section describes our experiment in more detail, followed by the results. All data and analysis files can be accessed at <https://osf.io/wxmeg/>.

2 The experiment

2.1 Participants

50 participants were recruited on Amazon’s Mechanical Turk. 20 participants were female, the remaining 30 male. Participants’ mean age was 39 (standard deviation: 11, range: 22–69). All participants indicated that they were native speakers of English. Participants were paid \$1.50 for their participation.

2.2 Materials

As mentioned above, the materials consisted of 16 adjectival scales. For each scale, we created a simple sentence containing the weaker scalar word. For each sentence, we created three types of images: one image where the sentence was unambiguously true, one image where it was unambiguously false, and one image where the sentence was true on its literal interpretation but false if its scalar inference was derived. [Table 3](#) shows the lexical scales that were tested; the Appendix provides the sentences and images used in the experiment.

The materials were pretested in two experiments with 25 participants each. Based on these pretests, we made several adjustments to the sentences and images to ensure that participants responded as expected, i.e., rejected the sentence in the false control condition, accepted it in the true control condition, and vacillated between accepting and rejecting the sentence in the target condition. (Here, vacillation was defined as significantly fewer ‘true’ responses than in the true control condition and significantly fewer ‘false’ responses than in the false control condition.)

The experiment presented each sentence-image pair three times, and thus comprised $16 \times 3 \times 3 = 144$ trials in total. The order of presentation was randomised for each participant.

2.3 Procedure

Participants were instructed to indicate whether or not the sentence was a good description of the image. They could register their judgement by pressing either ‘1’ (to answer in the positive) or ‘0’ (to answer in the negative) on their keyboard. Trials started with the presentation of the sentence. Upon pressing the space bar, the sentence was replaced by the image, whereupon participants could give their truth judgements. We measured the time from image onset to button press.

2.4 Data treatment

3 participants were removed from the analyses because their accuracy on control items was below 80%. In addition, we removed trials with a verification time faster than 200 milliseconds

or slower than 10 seconds, assuming that these correspond to accidental button presses or inattentiveness to the task at hand. This resulted in the removal of 14 trials (less than 0.1% of the data).

2.5 Results

Figure 1 shows the percentages of ‘true’ responses for each scalar word and condition. Performance in the control condition was close to ceiling. In the ‘true’ control condition, performance ranged from 87% for ‘warm’ to 100% for ‘sleepy’ and ‘content’. In the ‘false’ control condition, performance ranged from 86% for ‘unlikely’ to 100% for ‘ajar’ and ‘youthful’. In the target condition, the percentage of ‘true’ responses ranged from 16% for ‘mediocre’ to 80% for ‘youthful’.

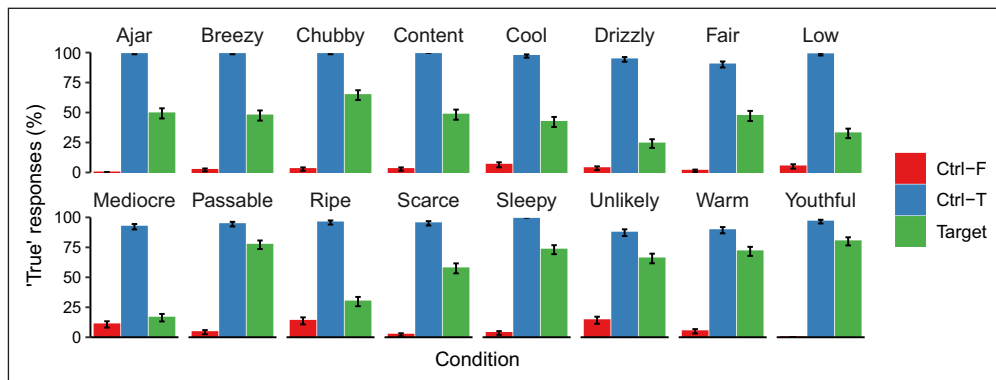


Figure 1 Percentage of ‘true’ responses for each scalar word and condition. Error bars represent standard errors of the mean.

One of our reviewers rightly observed that the percentage of ‘true’ responses in the target condition for ‘mediocre’ (16%) was so close to the percentage of errors in the ‘false’ control condition (11%) for that scalar word that one might call into question our assumption that the ‘not bad’ inference is a bona fide scalar inference rather than being an aspect of the lexical meaning of ‘mediocre’. However, given that ‘mediocre’ behaved in line with our assumption in the pretest (26% ‘true’ responses in the target condition vs. 7% errors in the ‘false’ control condition), we decided to retain this item in our analyses. We want to emphasise that removing ‘mediocre’ does not have any noteworthy consequences for the analyses that we report below.

Next, we considered participants’ verification times. **Figure 2** shows the mean logarithmised verification times for aggregated positive and negative scalar words. Here, we classified a scalar word as positive if its sign on the first principal component was positive; otherwise as negative (cf. **Table 3**). **Figure 2** shows the B&N effect for positive scalar words, but not for negative ones. Note that, as discussed previously, in the statistical analyses, we treated polarity as a continuous rather than binary factor; this visualisation is only intended to summarise our continuous polarity measure. **Figure 3** shows the mean logarithmised verification times for each scalar word separately.

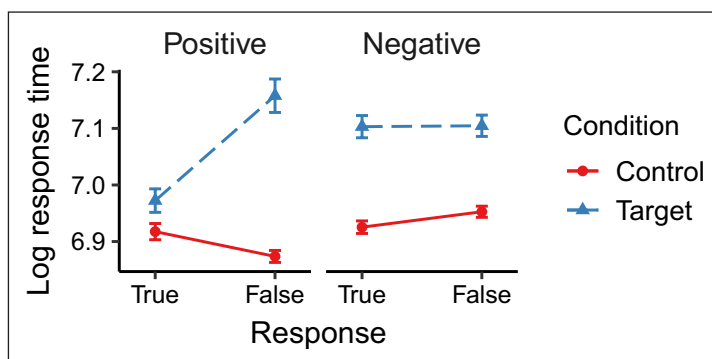


Figure 2 Mean logarithmised verification times for positive and negative scalar words. Error bars represent standard errors of the mean.

To test the polarity hypothesis, we constructed a linear regression mixed effects model predicting logarithmised response times based on condition (target vs. control), response (‘true’ vs. ‘false’), polarity, and their interactions, including random intercepts for participants and scalar words. For all analyses, degrees of freedom and corresponding *p*-values were estimated using the

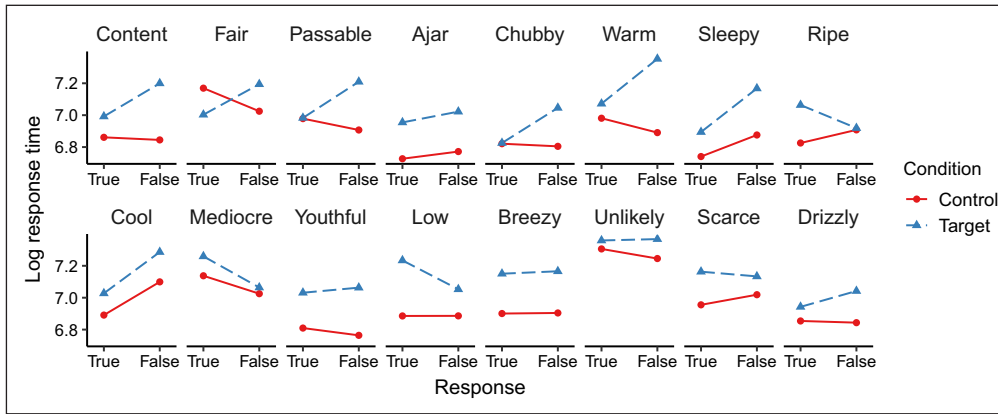


Figure 3 Mean logarithmised verification times for each scalar word. Scalar words are ordered from most positive (top left) to most negative (bottom right).

Satterthwaite procedure, as implemented in the ‘lmerTest’ package (Kuznetsova et al. 2013). The model also included response bias (i.e., the proportion of ‘true’ responses in the target condition) and trial number (linearly rescaled to the [0, 1] interval) as main effects. A response bias could lead to slower non-modal responses; trial number was included to capture learning effects. The model showed a highly significant interaction between condition, response, and polarity ($b = 0.09, SE = 0.01, t = 6.00, p < .001$). This interaction confirmed that polarity modulated the presence or absence of the B&N effect. There was no significant effect of response bias ($b = 0.02, SE = 0.02, t < 1$); however, there was a significant effect of trial number ($b = -0.50, SE = 0.02, t = 32.46, p < .001$) indicating that participants responded increasingly more rapidly throughout the experiment.

To also obtain a more fine-grained picture of the scope of the polarity hypothesis, we checked, for each scalar word separately, whether a B&N effect was present or not. To this end, for each scalar word, we constructed a linear regression mixed effects model predicting logarithmised response times on the basis of condition, response, their interaction, and trial number, including random intercepts for participants with random slopes for condition and response (but not their interaction). We observed significant interaction effects for ‘content’, ‘fair’, ‘passable’, ‘ajar’, ‘chubby’, ‘warm’, and ‘youthful’ (see *Table 4* for the full model parameters).

Scale	Polarity	Interaction effect			
		b	SE	t	p
(content, happy)	2.18	-0.35	0.10	-3.39	.001**
(fair, good)	2.12	-0.43	0.08	-5.14	.000***
(passable, good)	2.12	-0.30	0.09	-3.28	.001**
(ajar, open)	1.18	-0.19	0.08	-2.48	.014*
(chubby, fat)	1.06	-0.33	0.09	-3.85	.000***
(warm, hot)	0.46	-0.28	0.09	-3.14	.002**
(sleepy, asleep)	-0.03	-0.11	0.10	-1.20	.235
(ripe, overripe)	-0.28	0.19	0.10	1.96	.054
(cool, cold)	-0.36	-0.07	0.10	-0.73	.465
(mediocre, bad)	-0.69	-0.05	0.09	-0.57	.567
(youthful, young)	-0.85	-0.24	0.10	-2.33	.025*
(breezy, windy)	-0.99	-0.04	0.08	-0.51	.613
(drizzly, rainy)	-1.27	-0.10	0.08	-1.24	.216
(scarce, absent)	-1.31	0.08	0.08	0.94	.349
(low, empty)	-1.57	-0.13	0.08	-1.71	.089
(unlikely, impossible)	-1.83	0.05	0.09	0.56	.577

Table 4 Parameters of the interaction effect between condition (target vs. control) and response (‘true’ vs. ‘false’) for each lexical scale. The scales are ordered based on their estimated polarity value (*Polarity*, cf. *Table 3*).

Taken together, these results confirm the polarity hypothesis: the presence or absence of a B&N effect was modulated by the polarity of the scalar words. More specifically, all of the positive scalar words gave rise to a B&N effect, while almost none of the negative scalar words did. The only exception to this rule was 'youthful', which was associated with a B&N effect despite being assigned a (slightly) negative polarity value (-0.85).

3 General discussion

3.1 Summary

Pragmatic theories make conflicting predictions about the processing of scalar inferences in out-of-the-blue contexts. Relevance theory predicts that, in such contexts, the literal interpretation should be easier to retrieve than an interpretation that is enriched with a scalar inference. By contrast, Levinson (2000) predicts that it is the literal interpretation that should incur a processing cost, since it involves overturning the default enriched interpretation.

In a seminal study, Bott & Noveck (2004) found that the computation of the scalar inference of 'some' implying 'not all' was associated with increased sentence verification times. This *B&N effect* seems to provide strong evidence for the relevance-theoretic idea that the derivation of scalar inferences without a facilitating context is cognitively costly. However, more recent studies observed that the B&N effect does not consistently generalise beyond the ⟨some, all⟩ scale, which begs the question whether the B&N effect is really caused by the processing of the scalar inference (e.g., Chevallier et al. 2010; Romoli & Schwarz 2015; van Tiel et al. 2019b). To explain these findings, van Tiel and colleagues (2019b) hypothesised that the presence or absence of a B&N effect depends on the *polarity* of the scalar word.

The polarity-based explanation proceeds from the observation that verification times vary with the polarity of the sentence (e.g., Clark & Chase 1972; Carpenter & Just 1975). In particular, positive sentences are verified faster than sentences containing an implicitly negative word (e.g., 'low'), and the latter are verified faster than sentences containing an explicit negation (e.g., 'not all'). Correspondingly, given that all words on a scale have the same polarity, there are essentially three options: the words on a lexical scale may be positive, inherently negative, or explicitly negative. Hence, given that scalar inferences consist in the negation of the stronger scalar word, the polarity of the scalar inference is either negative (for positive scalar words) or doubly negative (for inherently or explicitly negative scalar expressions). Crucially, the polarity-based explanation argues that the latter are easier to verify than the former.

Various explanations for this potentially controversial assumption may be given. First, it could be that certain propositions containing a double negation are in fact easier to process than propositions with a single negation. Second, it could be that participants eliminate the double negation on the fly as they encounter it. Both of these explanations can be empirically tested by asking participants to verify the doubly negated scalar inferences (e.g., 'The battery is not empty') and comparing the verification times with those for the literal interpretation of the target sentence (i.e., 'The battery is low'). If either of the foregoing explanations is on the right track, 'false' responses to the negated scalar inference should be at least as fast as 'true' responses to the target sentence.

However, concerning the former explanation, prior research has shown that, in many cases, verification times increase with the number of negations (Sherman 1976). Hence, from a psychological perspective, perhaps the most plausible explanation is the latter: the positive form of the scalar inference is directly associated with its triggering expressions. According to this explanation, from a cognitive standpoint, the derivation of scalar inferences does not involve nonce construction and negation of alternatives, but rather resembles a form of disambiguation (cf. also Marty & Chemla 2013 for relevant comments about the parallelism between scalar inferencing and disambiguation).

If, finally, we assume that participants who arrive at a two-sided interpretation verify the literal interpretation and the scalar inference in parallel, the correct predictions follow straightforwardly: positive scalar words give rise to the B&N effect, inherently negative scalar words do not, and explicitly negated scalar words lead to the reverse B&N effect.

We extensively and systematically tested this polarity-based explanation by comparing the processing of 16 adjectival scalar inferences using a sentence-picture verification task. We estimated the polarity of the scalar words in our sample (and, hence, of the corresponding scalar inferences) on the basis of four diagnostics measuring their linguistic markedness and psychological valence. We found that the presence or absence of a B&N effect was strongly dependent on the polarity of the lexical scale. Indeed, of the 7 lexical scales whose inferences led to increased verification times, 6 were estimated to be positive. The sole exception was ⟨youthful, young⟩, which was associated with a B&N effect despite being classified as (somewhat) negative.

One interesting observation for this particular scale is that the valence criterion “correctly” classified this scale as positive rather than negative (i.e., in accordance with the behaviour of ‘youthful’ in the experiment, and in contrast to its estimated polarity). Hence, one may jump to the conclusion that the valence criterion generally offers a better measure of polarity than the other diagnostics. However, this does not hold true across the board. For example, ‘rainy’ also had positive valence ratings relative to ‘dry’, but the ⟨drizzly, rainy⟩ scale was not associated with a processing cost.

Another scalar word that merits some further discussion is ‘unlikely’. ‘Unlikely’ was the only scalar word in our sample that was explicitly marked for negativity by means of the negative prefix ‘un-’. Hence, one might expect to find a reverse B&N effect for this particular scalar word, i.e., one might expect that it patterns with explicitly negative scalar constructions like ‘not all’ rather than with implicitly negative scalar words like ‘low’. This prediction was not borne out. However, on closer inspection, this finding is not so surprising. To explain, consider the hierarchy of negation proposed by Fodor and colleagues (1975). According to Fodor and colleagues, negativity may have various sources. Ranging from “most negative” to “least negative”, these are as follows:

- i. Explicitly negative free morpheme (e.g., ‘not’).
- ii. Explicitly negative bound morpheme (e.g., ‘un-’).
- iii. Implicitly negative free morpheme (e.g., ‘low’).
- iv. Free morphemes that are defined in negative terms (e.g., ‘bachelor’ meaning someone who is *not* married, or ‘kill’ meaning causing someone to *not* be alive).

‘Unlikely’ is of class *ii*, whereas the other negative scalar words that we tested are all of class *iii*. Crucially, however, in terms of cognitive processing, Fodor and colleagues report that negative words of class *ii* pattern with implicitly negative words from class *iii*, rather than with the explicitly negative ones from class *i*.

Taken together, then, it is clear that there is a strong connection between polarity and the B&N effect. Perhaps most forcefully, it seems difficult to explain why the scalar inference of ‘warm’ but not ‘cool’ was associated with a processing cost without appealing to the notion of polarity, especially given that the sentences and images used for these scalar words were so similar (the images showed transparent drinking glasses containing water at different temperatures, from a block of ice to vigorously boiling, cf. Appendix). Hence, we view our results as strong support for the polarity-based explanation.

3.2 Processing scalar inferences

Crucially, if the polarity-based explanation is correct, the classic observation that certain scalar inferences lead to increased verification times is not reflective of any processing cost for scalar inferencing, but rather reflects the psychological difficulty of verifying negative information. Indeed, the polarity-based explanation leads us to conclude that scalar inferencing itself is not associated with a processing cost, even in the absence of a facilitating context, contra, e.g., relevance theory. At the same time, however, our results also fail to support the defaultist idea that scalar inferences arise temporally prior to the literal interpretation; nowhere did we observe faster processing times when people computed the scalar inference.

Rather, it seems that scalar inferences are conventionally or statistically associated with their triggering expressions. That is, when hearing an utterance containing the weaker scalar

word, people may immediately activate the corresponding scalar inference at no processing cost. However, when verifying this scalar inference, a processing cost may ensue if the scalar inference is negative, since negative information takes longer to be verified (e.g., Clark & Chase 1972; Carpenter & Just 1975).

One might suppose that this explanation is not very “Gricean” in spirit. Note, however, that Grice (1975) himself acknowledged the possibility that conversational implicatures are “intuitively grasped” (p. 50). What is crucial for Grice is whether or not this intuition is in principle “replaceable by an argument” (ibid.) that takes the literal meaning of the utterance and the assumption of cooperativity as its premises (see also Geurts & Rubio-Fernández 2015). However, such arguments should not necessarily be construed as psychologically real, i.e., hearers presumably do not actually construct such an argument every time they encounter a scalar word. Rather, the Gricean calculations provide a rational grounding of the inferences that hearers are entitled to derive (Kissine 2016; Geurts 2019; Dänzer 2020).

Our proposal makes some empirically testable predictions. In particular, it is predicted that other experimental paradigms that make use of sentence verification should also be susceptible to the polarity effect: if the proposition to be verified contains negative information, its verification should be cognitively costly. One such paradigm that relies on sentence verification was introduced by De Neys & Schaeken (2007). Their study essentially mirrored Bott and Noveck’s (2004: Exp. 3) in that participants gave intuitive truth judgements to underinformative sentences like (14). Crucially, however, De Neys and Schaeken required participants to memorise dot patterns of varying complexity during the process of sentence verification.

(14) Some dogs are mammals.

De Neys and Schaeken found that participants were less likely to respond ‘false’, i.e., to derive the scalar inference, when they had to memorise complex dot patterns compared to simple ones (cf. also De Neys & Schaeken 2007; Dieussaert et al. 2011; Marty & Chemla 2013; van Tiel et al. 2019a; Cho 2020; Marty et al. 2020). We refer to this finding as the *D&S effect*.

De Neys and Schaeken explain the D&S effect based on the premise that the derivation of scalar inferences is associated with a processing cost. According to their explanation, participants who had to memorise complex dot patterns had fewer cognitive resources available to derive the scalar inference, and consequently were less likely to carry out the derivation process. However, if the polarity-based explanation is on the right track, the D&S effect could also be modulated by the polarity of scalar words.

Interestingly, van Tiel and colleagues (2019b) also carried out a working memory load experiment for the same seven lexical scales that they tested in the sentence picture verification task that we discussed in the introduction. They found that, whereas all positive scalar words were associated with a D&S effect, the two negative scalar words in their sample—i.e., ‘low’ and ‘scarce’—were not. This observation suggests that the D&S effect is also susceptible to polarity in the same way as the B&N effect.

However, in a more recent study, Marty and colleagues (2020) provide provocative evidence against this conclusion by showing that the D&S effect is attested for indirect scalar inferences. For example, they found that, in displays showing only green apples, participants were less likely to reject sentences such as (15) when they had to memorise complex patterns than when they had to memorise simple ones.

(15) Not all of the apples are red.

To explain this pattern of results, Marty and colleagues (following Marty & Chemla 2013) argue that a processing cost emerges when participants make the cognitively costly decision to go beyond the literal interpretation. However, in line with the polarity-based explanation, they also hold that the process of deriving the scalar inference (i.e., the construction and rejection of alternatives) proceeds without any processing cost.

These findings paint a complex picture that obviously calls for a more detailed discussion than we can offer here, but they clearly show that, for other measures, too, there has been debate about whether the locus of the alleged processing cost is in the process of scalar inferencing or in some other relevant cognitive process (e.g., Marty & Chemla 2013; Politzer-Ahles & Husband

2018; Sun & Breheny 2019). This contribution fits into that line of work in showing that polarity is one of the factors influencing verification times in sentence verification tasks.

It is an open question whether polarity also influences other measures of processing cost, e.g., those involving reading times and eye movements. There is at least some evidence indicating that these measures, too, are influenced by the polarity of a sentence. For example, Glenberg et al. (1999) report longer reading times for negative sentences. Similarly, Tian et al. (2010) report that eye fixations to the correct image in a visual world paradigm are delayed for negative sentences compared to positive ones. However, both of these studies focus on sentences containing the explicit sentential negation ‘not’, rather than the implicitly negative words that we studied here. Hence, as it stands, it is unclear whether polarity has any explanatory role for experimental studies on the processing of scalar inferences using reading times and eye-tracking. Future work will have to determine whether polarity has pervasively influenced the experimental literature on scalar inference processing, or whether the effect is restricted to sentence verification as we have tacitly assumed throughout this paper.

3.3 Polarity

To estimate the polarity of the scalar words in our sample, we combined insights from linguistics and psychology. In linguistics, polarity is usually construed in terms of markedness; in psychology, in terms of subjective valence. We obtained measures of both construals, and used those to estimate a hypothesised latent construct of polarity. Here, we depart from (and hopefully improve on) prior research, particularly in linguistics.

Much linguistic research is premised on the idea that polarity is a binary notion, i.e., in any antonym pair, one is positive and the other one negative (e.g., Ruytenbeek et al. 2017; Gotzner et al. 2018). This approach regularly leads to an aporia. For example, Sassoon (2010) sought to determine polarity by looking at the frequency of antonym pairs in the ‘twice as [adjective]’ frame. In line with her hypothesis, intuitively positive adjectives tended to be more frequent in such frames than negative ones. However, there were ample exceptions. Thus, Sassoon found that ‘twice as bad’ was more frequent than ‘twice as good’, although she intuited that ‘good’ is positive and ‘bad’ is negative.

We observed many similar conflicts between diagnostics (cf. [Table 3](#)). For example, in Mohammad’s (2018) subjective valence study, ‘rainy’ was rated as more positive than ‘dry’, suggesting that ‘rainy’ is positive and ‘dry’ negative. However, the other diagnostics suggested the opposite conclusion. As one of our reviewers suggested, the “aberrant” outcome in terms of subjective valence could be due to the fact that ‘dry’ is highly polysemous, and that its subjective valence depends on which meaning is selected. For example, dry weather is generally considered positive, while dry bread is clearly negative. Hence, it could be that participants in Mohammad’s study tended to construe ‘dry’ in the latter negative sense rather than the former.

One way of resolving such clashes between diagnostics is by incorporating the results of multiple diagnostics into the estimation of a gradient measure of polarity. We hope this approach finds a following in linguistic and psychological research on polarity.

3.4 Conclusion

Perhaps above all, our results emphasise the importance of testing broader samples of scalar words in research on scalar inferences—not just to determine whether psychological effects generalise across the entire family of scalar words, but also because it offers an insight into the structural linguistic constructs that underlie language processing. Thus, our study has shown the psychological relevance of the notion of polarity. We hope our research will inspire others to revisit the many interesting findings that have been reported on the scalar inference of ‘some’ to see if they generalise and, if not, what factors may explain the observed scalar diversity, so that we may come to a better understanding of the cognitive processes that underlie the derivation of scalar inferences—and ultimately pragmatic inferences more generally.

Data Accessibility Statement

All data and analysis files can be found at <https://osf.io/wxmeq/>.

The additional file for this article can be found as follows:

- **Appendix.** The Appendix shows the sentences and images used in the experiment. DOI: <https://doi.org/10.5334/gjgl.1457.s1>

Acknowledgements

This research was presented at the workshop on degree expressions and polarity effects (DegPol2020) that was held at the Leibniz-Zentrum für Allgemeine Sprachwissenschaft. We thank the audience there for raising important questions and issues. We also thank Min-Joo Kim and our three anonymous reviewers at 'Glossa' for extremely valuable feedback on an earlier version of this article.

Funding information

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) and by the Dutch Science Organisation (Gravitation grant 'Language in Interaction', 024.001.006); both of which are gratefully acknowledged.

Competing Interests

The authors have no competing interests to declare.

Author affiliations

Bob van Tiel  orcid.org/0000-0002-4169-3179

Donders Institute for Brain, Cognition and Behaviour, Postbus 9010, 6500 GL Nijmegen, NL

Elizabeth Pankratz  orcid.org/0000-0001-8453-1105

Leibniz-Zentrum für Allgemeine Sprachwissenschaft, Schützenstraße 18, 10117 Berlin, DE

References

- Barbet, Cécile & Guillaume Thierry. 2018. When *some* triggers a scalar inference out of the blue. An electrophysiological study of a Stroop-like conflict elicited by single words. *Cognition* 177. 58–68. DOI: <https://doi.org/10.1016/j.cognition.2018.03.013>
- Benjafield, J. & J. Adams-Webber. 1976. The golden section hypothesis. *British Journal of Psychology* 67. 11–15. DOI: <https://doi.org/10.1111/j.2044-8295.1976.tb01492.x>
- Bott, Lewis & Ira A. Noveck. 2004. Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language* 51. 437–457. DOI: <https://doi.org/10.1016/j.jml.2004.05.006>
- Boucher, Jerry & Charles E. Osgood. 1969. The Pollyanna hypothesis. *Journal of Verbal Learning and Verbal Behavior* 8. 1–8. DOI: [https://doi.org/10.1016/S0022-5371\(69\)80002-2](https://doi.org/10.1016/S0022-5371(69)80002-2)
- Breheny, Richard, Napoleon Katsos & John Williams. 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: <https://doi.org/10.1016/j.cognition.2005.07.003>
- Carpenter, Patricia A. & Marcel A. Just. 1975. Sentence comprehension: A psycholinguistic model of sentence verification. *Psychological Review* 82. 45–73. DOI: <https://doi.org/10.1037/h0076248>
- Chemla, Emmanuel & Lewis Bott. 2014. Processing inferences at the semantics/pragmatics frontier: Disjunctions and free choice. *Cognition* 130. 380–396. DOI: <https://doi.org/10.1016/j.cognition.2013.11.013>
- Cheng, Chao-Ming & Huei-Jane Huang. 1980. The process of verifying affirmative and negative sentences against pictures. *Memory & Cognition* 8. 573–583. DOI: <https://doi.org/10.3758/BF03213777>
- Chevallier, Coralie, Deirdre Wilson, Francesca Happé & Ira Noveck. 2010. Scalar inferences in autism spectrum disorders. *Journal of Autism and Developmental Disorders* 40. 1104–1117. DOI: <https://doi.org/10.1007/s10803-010-0960-8>
- Chevallier, Coralie, Ira A. Noveck, Tatjana Nazir, Lewis Bott, Valentina Lanzetti & Dan Sperber. 2008. Making disjunctions exclusive. *Quarterly Journal of Experimental Psychology* 61. 1741–1760. DOI: <https://doi.org/10.1080/17470210701712960>
- Cho, Jacee. 2020. Memory load effect in the real-time processing of scalar implicature. *Journal of Psycholinguistic Research* 49. 865–884. DOI: <https://doi.org/10.1007/s10936-020-09726-3>

- Clark, Herbert H. 1974. The chronometric study of meaning components. In Jacques Mehler (ed.), *Problems actuels en psycholinguistique*, 490–505. Paris, France: Centre National de la Recherche Scientifique.
- Clark, Herbert H. & Eve V. Clark. 1977. *Psychology and language: An introduction to psycholinguistics*. New York, NY: Harcourt.
- Clark, Herbert H. & William G. Chase. 1972. On the process of comparing sentences against pictures. *Cognitive Psychology* 3. 472–517. DOI: [https://doi.org/10.1016/0010-0285\(72\)90019-9](https://doi.org/10.1016/0010-0285(72)90019-9)
- Cremers, Alexandre & Emmanuel Chemla. 2014. Direct and indirect scalar implicatures share the same processing signature. In Salvatore Pistoia Reda (ed.), *Pragmatics, semantics and the case of scalar implicatures*, 201–227. London, United Kingdom: Palgrave Macmillan. DOI: https://doi.org/10.1057/9781137333285_8
- Dänzer, Lars. 2020. The explanatory project of Gricean pragmatics. *Mind & Language*. DOI: <https://doi.org/10.1111/mila.12295>
- De Neys, Wim & Walter Schaeken. 2007. When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology* 54. 128–133. DOI: <https://doi.org/10.1027/1618-3169.54.2.128>
- Degen, Judith & Michael K. Tanenhaus. 2016. Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science* 40. 172–201. DOI: <https://doi.org/10.1111/cogs.12227>
- Degen, Judith & Michael K. Tanenhaus. 2019. Constraint-based pragmatic processing. In C. Cummins & N. Katsos (eds.), *Handbook of experimental pragmatics*, 21–38. Oxford, United Kingdom: University Press. DOI: <https://doi.org/10.1093/oxfordhb/9780198791768.013.8>
- Diussaert, Kristien, Suzanne Verkerk, Ellen Gillard & Walter Schaeken. 2011. Some effort for some: Further evidence that scalar implicatures are effortful. *Quarterly Journal of Experimental Psychology* 64. 2352–2367. DOI: <https://doi.org/10.1080/17470218.2011.588799>
- Fauconnier, Gilles. 1975. Pragmatic scales and logical structure. *Linguistic Inquiry* 6. 353–375.
- Fodor, Janet D., Jerry A. Fodor & Merrill F. Garrett. 1975. The psychological unreality of semantic representations. *Linguistic Inquiry* 6. 515–531.
- Gazdar, Gerald. 1979. *Pragmatics: implicature, presupposition, and logical form*. New York, NY: Academic Press.
- Geurts, Bart. 2010. *Quantity implicatures*. Cambridge, United Kingdom: University Press. DOI: <https://doi.org/10.1017/CBO9780511975158>
- Geurts, Bart. 2019. Communication as commitment sharing: Speech acts, implicatures, common ground. *Theoretical Linguistics* 45. 1–30. DOI: <https://doi.org/10.1515/tl-2019-0001>
- Geurts, Bart & Paula Rubio-Fernández. 2015. Pragmatics and processing. *Ratio* 28. 446–469. DOI: <https://doi.org/10.1111/rati.12113>
- Givón, Talmy. 1979. *On understanding grammar*. New York, NY: Academic Press.
- Glenberg, Arthur M., David A. Robertson, Jennifer L. Jansen & Mina C. Johnson-Glenberg. 1999. Not propositions. *Journal of Cognitive Systems Research* 1. 19–33. DOI: [https://doi.org/10.1016/S1389-0417\(99\)00004-2](https://doi.org/10.1016/S1389-0417(99)00004-2)
- Gotzner, Nicole, Stephanie Solt & Anton Benz. 2018. Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology* 9. 1659. DOI: <https://doi.org/10.3389/fpsyg.2018.01659>
- Greenberg, Joseph H. 1966. *Language universals, with special reference to feature hierarchies*. The Hague, The Netherlands: Mouton.
- Grice, H. P. 1975. Logic and conversation. In Peter Cole & Jerry L. Morgan (eds.), *Syntax and semantics, volume 3: Speech acts*, 41–58. New York, NY: Academic Press.
- Grodner, Daniel J., Natalie M. Klein, Kathleen M. Carbary & Michael K. Tanenhaus. 2010. “Some,” and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition* 116. 42–55. DOI: <https://doi.org/10.1016/j.cognition.2010.03.014>
- Heim, Irene. 2008. Decomposing antonyms? In A. Grønn (ed.), *Proceedings of Sinn und Bedeutung 12*, 212–225. Oslo, Norway: ILOS.
- Horn, Laurence R. 1972. *On the semantic properties of logical operators in English*. University of California, Los Angeles dissertation.
- Horn, Laurence R. 1989. *A natural history of negation*. Chicago, IL: University Press.
- Huang, Yan. 2014. *Pragmatics*. Oxford, United Kingdom: University Press.
- Huang, Yi Ting & Jesse Snedeker. 2018. Some inferences still take time: Prosody, predictability, and the speed of scalar implicatures. *Cognitive Psychology* 102. 105–126. DOI: <https://doi.org/10.1016/j.cogpsych.2018.01.004>
- Ingram, Joanne, Christopher J. Hand & Greg Maciejewski. 2016. Exploring the measurement of markedness and its relationship with other linguistic variables. *PLOS ONE* 11. e0157141. DOI: <https://doi.org/10.1371/journal.pone.0157141>
- Kennedy, Chris & Louise McNally. 2005. Scale structure, degree modification, and the semantics of gradable predicates. *Language* 81. 345–381. DOI: <https://doi.org/10.1353/lan.2005.0071>

- Kissine, Mikhail. 2016. Pragmatics as metacognitive control. *Frontiers in Psychology* 6. 2057. DOI: <https://doi.org/10.3389/fpsyg.2015.02057>
- Kuznetsova, Alexandra, Per Bruun Brockhoff & Rune Haubo Bojensen Christensen. 2013. lmerTest: Tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package) [R package].
- Lehrer, Adrienne. 1985. Markedness and antonymy. *Journal of Linguistics* 21. 397–429. DOI: <https://doi.org/10.1017/S002222670001032X>
- Lehrer, Adrienne & Keith Lehrer. 1982. Antonymy. *Linguistics and Philosophy* 5. 483–501. DOI: <https://doi.org/10.1007/BF00355584>
- Levinson, Stephen. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press. DOI: <https://doi.org/10.7551/mitpress/5526.001.0001>
- Lyons, John. 1968. *Introduction to theoretical linguistics*. Cambridge, United Kingdom: University Press.
- Marty, Paul & Emmanuel Chemla. 2013. Scalar implicatures: Working memory and a comparison with only. *Frontiers in Psychology* 4. 1–12. DOI: <https://doi.org/10.3389/fpsyg.2013.00403>
- Marty, Paul, Jacopo Romoli, Yasutada Sudo, Bob van Tiel & Richard Breheny. 2020. Processing implicatures: A comparison between direct and indirect sis. Oral presentation at Experiments in Linguistic Meaning (ELM), Philadelphia, PA, September 16–18, 2020.
- Mohammad, Saif M. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 english words. In *Proceedings of the annual conference of the association for computational linguistics (ACL)*. Melbourne, Australia. DOI: <https://doi.org/10.18653/v1/P18-1017>
- Moracchini, Sophie. 2019. *Morphosyntax and semantics of degree constructions*. Massachusetts Institute of Technology, Boston, MA dissertation.
- Morzycki, Marcin. 2015. *Modification*. Cambridge, United Kingdom: University Press. DOI: <https://doi.org/10.1017/CBO9780511842184>
- Nieuwland, Mante S. & Gina R. Kuperberg. 2008. When the truth is not too hard to handle. *Psychological Science* 19. 1213–1218. DOI: <https://doi.org/10.1111/j.1467-9280.2008.02226.x>
- Nouwen, Rick. 2020. Evaluation, extent, and Goldilocks. Unpublished manuscript, Utrecht University, The Netherlands.
- Noveck, Ira A. & A. Posada. 2003. Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language* 85. 203–210. DOI: [https://doi.org/10.1016/S0093-934X\(03\)00053-1](https://doi.org/10.1016/S0093-934X(03)00053-1)
- Noveck, Ira A. & Dan Sperber. 2007. The why and how of experimental pragmatics: The case of ‘scalar inferences’. In N. Burton-Roberts (ed.), *Advances in pragmatics*, 184–212. Basingstoke, United Kingdom: Palgrave. DOI: https://doi.org/10.1057/978-1-349-73908-0_10
- Osgood, Charles & Meredith Martin Richards. 1973. From Yang and Yin to *and* or *but*. *Language* 49. 380–412. DOI: <https://doi.org/10.2307/412460>
- Paradis, Carita, Joost van de Weijer, Caroline Willners & Magnus Lindgren. 2012. Evaluative polarity of antonyms. *Lingue e Linguaggio* 11. 199–214.
- Politzer-Ahles, Stephen & Matthew E. Husband. 2018. Eye movement evidence for context-sensitive derivation of scalar inferences. *Collabra* 1. 1–13. DOI: <https://doi.org/10.1525/collabra.100>
- Proctor, Robert W. & Yang Seok Cho. 2006. Polarity correspondence: A general principle for performance of speeded binary classification tasks. *Psychological Bulletin* 132. 416–442. DOI: <https://doi.org/10.1037/0033-2909.132.3.416>
- Récanati, François. 1995. The alleged priority of literal interpretation. *Cognitive Science* 19. 207–232. DOI: https://doi.org/10.1207/s15516709cog1902_2
- Rett, Jessica. 2008. *The semantics of evaluativity*. Oxford, United Kingdom: University Press.
- Romoli, Jacopo & Florian Schwarz. 2015. An experimental comparison between presuppositions and indirect scalar implicatures. In F. Schwarz (ed.), *Experimental perspectives on presuppositions*, 215–240. Cham, Germany: Springer. DOI: https://doi.org/10.1007/978-3-319-07980-6_10
- Ronai, Eszter & Ming Xiang. 2020. Pragmatic inferences are QUD sensitive: An experimental study. *Journal of Linguistics*. DOI: <https://doi.org/10.1017/S0022226720000389>
- Ruytenbeek, Nicolas, Steven Verheyen & Benjamin Spector. 2017. Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives. *Glossa* 2. 1–27. DOI: <https://doi.org/10.5334/gjgl.151>
- Sassoon, Galit W. 2010. The degree functions of negative adjectives. *Natural Language Semantics* 18. 141–181. DOI: <https://doi.org/10.1007/s11050-009-9052-8>
- Schäfer, Roland. 2015. Processing and querying large web corpora with the COW14 architecture. In Piotr Bański, Hanno Biber, Evelyn Breiteneder, Marc Kupietz, Harald Lungen & Andreas Witt (eds.), *Proceedings of challenges in the management of large corpora 3 (CMC-3)*, 28–34. Lancaster, United Kingdom: IDS.
- Schäfer, Roland & Felix Bildhauer. 2012. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk & Stelios Piperidis (eds.), *Proceedings of the eighth international conference on language resources and evaluation (LREC'12)*, 486–493. Istanbul, Turkey: ELRA.

- Sherman, Mark A. 1976. Adjectival negation and the comprehension of multiply negated sentences. *Journal of Verbal Learning and Verbal Behavior* 15. 143–157. DOI: [https://doi.org/10.1016/0022-5371\(76\)90015-3](https://doi.org/10.1016/0022-5371(76)90015-3)
- Soames, Scott. 1982. How presuppositions are inherited: A solution to the projection problem. *Linguistic Inquiry* 13. 483–545.
- Solt, Stephanie. 2015. Measurement scales in natural language. *Language and Linguistics Compass* 9. 14–32. DOI: <https://doi.org/10.1111/lnc3.12101>
- Sperber, Dan & Deirdre Wilson. 1987. Précis of relevance: communication and cognition. *Behavioral and Brain Sciences* 10. 697–754. DOI: <https://doi.org/10.1017/S0140525X00055345>
- Sperber, Dan & Deirdre Wilson. 1995. *Relevance: communication and cognition* (2nd edition). Oxford, United Kingdom: Blackwell.
- Sun, Chao & Richard Breheny. 2019. Another look at the online processing of scalar inferences: An investigation of conflicting findings from visual-world eye-tracking studies. *Language, Cognition and Neuroscience*. DOI: <https://doi.org/10.1080/23273798.2019.1678759>
- Tian, Ye, Richard Breheny & Heather J. Ferguson. 2010. Why we simulate negated information: A dynamic pragmatic account. *Quarterly Journal of Experimental Psychology* 63. 2305–2312. DOI: <https://doi.org/10.1080/17470218.2010.525712>
- Tomlinson Jr., John M., Todd M. Bailey & Lewis Bott. 2013. Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language* 69. 18–35. DOI: <https://doi.org/10.1016/j.jml.2013.02.003>
- van Tiel, Bob, Elizabeth Pankratz & Chao Sun. 2019b. Scales and scalarity: Processing scalar inferences. *Journal of Memory and Language* 105. 427–441. DOI: <https://doi.org/10.1016/j.jml.2018.12.002>
- van Tiel, Bob, Elizabeth Pankratz, Paul Marty & Chao Sun. 2019a. Scalar inferences and cognitive load. In M. Teresa Espinal, E. Castroviejo, M. Leonetti, L. McNally & C. Real-Puigdollers (eds.), *Proceedings of Sinn und Bedeutung* 23, 429–443. Bellaterra, Spain: Universitat Autònoma de Barcelona.
- van Tiel, Bob, Emiel van Miltenburg, Natalia Zevakhina & Bart Geurts. 2016. Scalar diversity. *Journal of Semantics* 33. 107–135. DOI: <https://doi.org/10.1093/jos/ffu017>
- van Tiel, Bob, Michael Franke & Uli Sauerland. 2021. Probabilistic pragmatics explains gradience and focality in natural language quantification. *Proceedings of the National Academy of Sciences of the United State of America* 118. e200545311. DOI: <https://doi.org/10.1073/pnas.2005453118>
- van Tiel, Bob & Walter Schaeken. 2016. Processing conversational implicatures: Alternatives and counterfactual reasoning. *Cognitive Science* 41. 1–36. DOI: <https://doi.org/10.1111/cogs.12362>
- Wason, P. C. 1959. The processing of positive and negative information. *Quarterly Journal of Experimental Psychology* 11. 92–107. DOI: <https://doi.org/10.1080/17470215908416296>
- Wason, P. C. 1965. The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior* 4. 7–11. DOI: [https://doi.org/10.1016/S0022-5371\(65\)80060-3](https://doi.org/10.1016/S0022-5371(65)80060-3)
- Wason, P. C. 1972. In real life negatives are false. *Logique et Analyse* 15. 17–38.
- Westera, Matthijs. 2017. *Exhaustivity and intonation: A unified theory*: University of Amsterdam, The Netherlands dissertation.

van Tiel and Pankratz
*Glossa: a journal of
 general linguistics*
 DOI: 10.5334/gjgl.1457

TO CITE THIS ARTICLE:

van Tiel, Bob and Elizabeth Pankratz. 2021. Adjectival polarity and the processing of scalar inferences. *Glossa: a journal of general linguistics* 6(1): 32. 1–21. DOI: <https://doi.org/10.5334/gjgl.1457>

Submitted: 02 October 2020

Accepted: 07 February 2021

Published: 31 March 2021

COPYRIGHT:

© 2021 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

Glossa: a journal of general linguistics is a peer-reviewed open access journal published by Ubiquity Press.