Journal of Memory and Language 69 (2013) 18–35

Contents lists available at SciVerse ScienceDirect

Journal of Memory and Language

journal homepage: www.elsevier.com/locate/jml

Possibly all of that and then some: Scalar implicatures are understood in two steps

John M. Tomlinson Jr. [a,*], Todd M. Bailey [b], Lewis Bott [b]

[a] Zentrum für Allegemeine Sprachwissenschaft (ZAS), Germany
[b] School of Psychology, Cardiff University, United Kingdom

ARTICLE INFO

Article history:
Received 7 May 2012
revision received 19 February 2013
Available online 26 March 2013

Keywords:
Pragmatics
Scalar implicatures
Inference
Psycholinguistics
Mouse-tracking

ABSTRACT

Scalar implicatures often incur a processing cost in sentence comprehension tasks. We used a novel mouse-tracking technique in a sentence verification paradigm to test different accounts of this effect. We compared a two-step account, in which people access a basic meaning and then enrich the basic meaning to form the scalar implicature, against a one-step account, in which the scalar implicature is directly incorporated into the sentence representation. Participants read sentences and used a computer mouse to indicate whether each sentence was true or false. Three experiments found that when verifying sentences like "some elephants are mammals", average mouse paths initially moved towards the true target and then changed direction mid-flight to select the false target. This supports the two-step account of implicatures. We discuss the results in relation to previous findings on scalar implicatures and theoretical accounts of pragmatic inference.

Introduction

To communicate efficiently, speakers often imply information instead of explicitly stating it. Consider this exchange:

(1A) Nowadays, teenagers are tethered to their smart phones.
(1B) Some are.

Here, B is a teenager who distances himself from people of his age who seemingly never put down their mobile phones. By saying, "some are," he confirms that there are indeed teenagers who match A's description. More importantly for the purposes of this paper, he also implies that there is a significant group of teenagers who do not use their phones excessively.

In order to understand inferences like those above, the listener must know which of an infinite number of potential inferences the speaker intended her to draw. Moreover, for the sake of efficiency and communicative fluency, the inferences must be derived in a very short space of time. Grice's (1975, 1989) maxims of communication describe abstract principles that could guide the listener in drawing inferences. However, something like Grice's maxims might be realized by any number of processing mechanisms. In this paper, we test between two processing models of scalar implicatures (see also, Bott & Noveck, 2004; Breheny, Katsos, & Williams, 2006; Huang & Snedeker, 2009). The first model assumes the listener derives the implicature in a single processing step – a one-step model – and the second assumes the listener initially derives a literal, or basic, meaning, and then enriches this to form the implicature – a two step model.

The structure of the paper is as follows. We first introduce scalar implicatures in more detail and present a summary of the relevant linguistic literature. We then present the two processing models in more detail and discuss how they account for previous findings on processing scalar

* Corresponding author. Address: Zentrum für Allegemeine Sprachwissenschaft (ZAS), Schützenstr. 18, 10117 Berlin, Germany. Fax: +49 (0)30 20192 402.
E-mail address: tomlinson@zas-gwz-berlin.de (J.M. Tomlinson Jr.).

0749-596X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jml.2013.02.003

implicatures. Finally, we introduce the paradigm that we use to test between the models and describe three experiments that test the model predictions.

Scalar implicatures

The inference in (1) is an example from a broader group of inferences known as *scalar implicatures* (see Geurts, 2010, for a thorough discussion). When B says "Some are," in (1) he implies that *not all* teenagers use their phones excessively. This inference can be described using Grice's (1975) Cooperative Principle and general reasoning abilities. According to the Gricean explanation, the listener first computes some sort of basic meaning for what was said (e.g. "at least some..."). This is contrasted with more informative and relevant things that the listener could have said instead, if they had been true. For example, in (1B) the speaker said, "some are," but he could have said, "all are", which would have been more informative and relevant. Relying on the Cooperative Principle, the listener assumes that the speaker would have used the more informative statement if it were true. Because the speaker did not, the listener infers that *all* must not hold. Finally, by combining what the speaker said, "some are," with the *not all* inference, the listener arrives at the final interpretation, *some but not all are*.

In general, scalar implicatures occur when a speaker uses a weak element from a scale of elements ordered in terms of semantic strength (a *semantic* or *Horn* scale; see Horn, 1972, 1989). Under these circumstances the listener is licensed to infer that the stronger elements in the scale do not hold. For example, *some*, *many*, *all*, form a semantic scale, *some* < *many* < *all*, with *all* being the strongest, most informative element (whenever *all X* is true, *some X* and *many X* are also true, but not the reverse). Use of *some* can therefore imply the negation of *many* and *all*. Other examples of semantic scales and their associated implicatures include, *may* < *must*, where the use of *may* can imply *not must*; *or* < *and*, where *or* can imply *not and*, and *warm* < *hot*, where *warm* implies *not hot*. Indeed, any set of elements can become part of a semantic scale and generate scalar implicatures in a suitable context, as in the scale, *handsome* < *handsome and intelligent*, that arises from speaker A saying, "John's handsome and intelligent" and speaker B responding with, "Well, he's handsome," (see Carston, 1998). As with other pragmatic phenomena, scalar implicatures are defeasible, or cancellable (e.g., "some are... in fact all of them are."). Defeasibility distinguishes scalar implicatures from entailments, but unlike other pragmatic phenomena, scalar implicatures often occur in very structured semantic environments (see e.g., Chierchia, 2004). For example, scalar implicatures do not arise when used in the antecedent of the conditional ("If some of the children are in the classroom,..."), and they interact systematically with negation, such as the *some* implication that arises when a speaker says *not all*, as in "Not all of the children are in the classroom." Thus scalar implicatures involve interactions between semantic and pragmatic considerations, providing a unique domain in which to employ insights from two often separate disciplines of study (see Horn, 2006, "The border wars").

In psycholinguistic investigations of how scalar implicatures are processed, most work has considered a processing adaptation of Neo-Gricean theory (e.g., Gazdar, 1979; Levinson, 2000), known as the *default model*. According to Levinson, for example, quantificational determiners such as *some* are associated with alternative constructions in memory (e.g., *all* and *many*). The contrast between the expression used (e.g. "some") and an alternative construction that was not used automatically leads to the implicature (e.g., *not all*). In the processing literature this has been taken to mean that scalar implicatures should arise on every occasion in which a scalar term occurs, but that subsequently the implicature is sometimes cancelled (e.g., Bott & Noveck, 2004; Breheny et al., 2006; Huang & Snedeker, 2009). In other words, the implicature arises by default. Although this work is important, and we discuss it in more detail below, our approach to processing of scalar implicatures takes a different tack. Instead of asking whether the implicature is derived by default even when it is not required, we ask how that derivation takes place: does deriving an implicature involve a single processing step, or are there multiple steps?

One-step versus two-step processing models

We suggest a distinction between, on the one hand, computing a basic meaning and then enriching it to form a different meaning, and on the other, computing the completed meaning in a single processing step. We refer to the former as two-step models, and the latter as one-step models.

Two-step models are those in which an initial semantic interpretation forms a basis from which a distinctly different meaning is eventually derived. Several different theories are possible; the most obvious being a processing version of a Gricean account. Under this view, a listener must first compute the literal meaning of the sentence and its possible alternatives (Step 1), and then, assuming the speaker is informative and reliable, the listener enriches the literal meaning with the implicature (Step 2). The output of Step 1 is necessary to execute Step 2. Alternatively, the default implicature model (as described above), in which the implicature is always derived but sometimes cancelled, is also an example of a two-step model, albeit with Step 1 corresponding to the implicature and Step 2, after cancelling, corresponding to the literal meaning. Other examples include a model in which the decision to proceed onto Step 2 processing is not contingent on the output of Step 1 but nonetheless automatically follows it. The common theme running through two-step models is that some form of meaning is used as a basis to derive a different, second meaning.

One-step models, on the other hand, do not assume multiple, sequential processing steps. Necessary computations can be made in parallel and the appropriate scalar interpretation can be incorporated into the sentence in a single processing step, rather like constraint-based models of processing in which contextual, grammatical and other factors are all computed in parallel to provide the best guess at the appropriate interpretation (e.g., Bates & MacWhinney, 1989; MacDonald, Pearlmutter, & Seidenberg, 1994; van

Gompel, Pickering, Pearson, & Liversedge, 2005; Degen & Tanenhaus, 2011, who explicitly propose a model of this sort for scalar implicatures). One-step models can be conceptualized by assuming that each scalar interpretation (the literal meaning and the implicature) has an activation level, determined by the context, and the interpretation with the highest activation is incorporated into the sentence. For example, consider the difference between *some* and *some of*. The increased likelihood of an implicature interpretation for *some of* compared to *some* (Grodner, Klein, Carbary, & Tanenhaus, 2010) could be explained by supposing that the presence of the partitive *of* increases activation levels of the implicature interpretation. Conversely, expressions that appear to block scalar implicatures, such as *if* (Chierchia, 2004) or *not* (Gazdar, 1979), could decrease the activation levels. In principle, there are many ways that implicatures could be derived by different one-step models. Scalar interpretations might be stored directly in the lexicon, for example, or procedures for deriving the upper-bound interpretation[1] might be precompiled and triggered when necessary (along the lines of Chierchia, 2004, 2006). The common theme in one-step models is that multiple interpretations are not accessed in sequence: a single meaning is selected, either upper- or lower-bound, and incorporated directly into the sentence representation.

The work presented in this paper tests between these two accounts. Before describing our experiments, however, we review the previous literature on scalar implicature processing and illustrate how the models above account for those findings.

Processing scalar implicatures

A growing body of evidence suggests that sentences with scalar implicatures incur processing costs. Upper-bound interpretations, such as *some* [but not all], have long response latencies relative to lower-bound interpretations, such as *some* [and possibly all], in sentence verification tasks (Bott & Noveck, 2004), and longer reading times (Bergen & Grodner, 2012; Breheny et al., 2006). They show delayed onset of above-chance accuracy in sentence verification tasks with response deadlines (Bott, Bailey, & Grodner, 2012), and delayed eye fixations in a visual world task (Huang & Snedeker, 2009; but see Grodner et al., 2010). As we shall see, one-step and two-step models both account for these findings but in different ways. In our review of the literature, we start with a discussion of Bott and Noveck, whose methodology forms the basis of this study, and we also consider the visual world studies, which have found mixed evidence for the cost of implicatures.

Bott and Noveck (2004) argued that according to the default model, the time needed to derive a scalar implicature will always be less than the time needed to derive the literal meaning (see also Noveck & Posada, 2003). They trained participants to respond either true or false to underinformative sentences such as *some elephants are mammals*. These sentences were ambiguous in that they were true if participants derived the lower-bound meaning of the scalar term, as in, *some* [and possibly all] *elephants are mammals*, but false if participants derived the upper-bound meaning, as in, *some* [but not all] *elephants are mammals*. Bott and Noveck found that upper-bound interpretations (false) were slower than lower-bound interpretations (true) for the experimental sentences, yet there was very little difference between comparable true and false control sentences. Indeed, lower-bound interpretations were processed at the same speed as true control sentences. In another experiment, Bott and Noveck reversed the mapping between the true and false response options and the upper and lower-bound interpretations, and found that the upper-bound interpretations were slow even when upper-bound interpretations were associated with true responses. Bott and Noveck therefore concluded that there was a cost to deriving scalar implicatures that did not apply to interpretations without the implicatures, and that, consequently, scalar implicatures were not computed by default.

Bott et al. (2012) adapted Bott and Noveck's (2004) sentence verification task, using a response deadline procedure to eliminate the possibility of speed–accuracy trade-off effects (see McElree, 1993; Reed, 1973). More processing time was required to achieve above-chance accuracy for upper-bound interpretations than for lower-bound interpretations, consistent with Bott and Noveck's study. Bott et al. also found that participants were faster to correctly respond to *only some* sentences (*only some elephants are mammals*) than to equivalent upper-bound *some* (*some* [but not all] *elephants are mammals*). They concluded that implicatures incur additional costs that are likely to reflect some form of pragmatic enrichment applied to a prior literal meaning to derive the upper-bound interpretation. An interesting alternative explanation, however, concerns the focusing properties of *only*. If *only* focuses attention on the complement set (e.g., elephants who are not mammals), whereas upper-bound *some* [but not all] focuses on the referent set (elephants that are mammals), participants in Bott et al. (2012) might have found it easier to reject the *only* statements than the plain *some* statements. Some intuitive evidence for this is given by consideration of sentences like *only some elephants are cars*, which are false, but appear more difficult to reject than sentences like *some elephants are cars*, which are also false. The focusing properties of *only* might explain this difference. In short, while Bott et al., present suggestive evidence that part of the cost observed by Bott and Noveck (2004) was due to pragmatic enrichment, there might be other, non-pragmatic explanations for the Bott et al. *only* results.

Sentence verification tasks have found a cost to scalar implicatures, as have sentence reading studies (Bergen & Grodner, 2012; Breheny et al., 2006). Visual world studies present a more varied picture, however. Consistent with sentence verification results, Huang and Snedeker (2009) found that looking times to a referent were delayed when participants needed to form a scalar implicature, relative to

[1] We refer to scalar sentences as having *upper-bound* meanings when they have *some but not all* interpretations, and *lower-bound* meanings when they have the literal meaning, or *some and possibly all* interpretations, consistent with Breheney et al. (2006). The terminology refers to the scale having a bounded meaning at the upper end of the semantic scale in the implicature case (something less than *all*).

a control quantifier. Participants saw a display of four images and heard sentences instructing them to click on one of them. The images consisted of characters with a variety of objects. In the critical trials, one of the characters (e.g., a girl) had a subset of one item (e.g., some of the socks), another character (a boy) had a subset of those same items (socks), and third and fourth characters had sets of other items. The participant heard sentences like, "Click on the girl with [quantifier] of the socks." The quantifier was either the scalar term, *some*, or a control quantifier, such as *two* or *all*. Huang and Snedeker found that eye fixations to the target image were delayed when the quantifier was upper-bound *some* compared to the control quantifiers. They argued that this delay was because the participant needed to compute an implicature in the *some* case, but not for the other quantifiers.

A similar visual world study, however, by Grodner et al. (2010), failed to observe delayed looking times for upper-bound *some*. Grodner et al. suggested several differences in procedure that might explain the discrepancy. For example, they equated target image size across conditions so that delays in the *some* condition could not be explained by preferential looking to the larger image (in Huang and Snedeker's, 2009, study, the *all* conditions involved larger images). One particularly important difference was the presence of more apt amount descriptors (*two* and *three*) in the filler items of Huang and Snedeker, but not in Grodner et al. Participants in Huang and Snedeker's study might therefore have encoded the target images using the numerical quantifiers, such as, *the girl with two of the socks*, and then experienced difficulty relating the target sentence, *the girl with some of the socks*, back to their encoded image. The delay observed by Huang and Snedeker might therefore be explained by a mismatch between the way in which images were encoded and the target sentence (see Degen & Tanenhaus, 2011 and Grodner et al.). Grodner et al. concluded that there might be contexts in which there is no cost to scalar implicatures, or that the cost in referential paradigms is much smaller than previously thought (100 ms or less).

The studies reviewed above suggest that interpreting a sentence with a scalar implicature carries a cost, on at least some occasions. The studies that directly compared upper-bound with lower-bound interpretations all found that upper-bound interpretations required more processing time than the lower-bound interpretation. Comparisons between upper-bound *some* and other quantifiers have produced mixed results, but none of the studies found an advantage in processing time for upper-bound interpretations.

The one-step and the two-step models described above can both explain costs associated with scalar implicatures, suggesting that interpretation time is not sufficient to distinguish between them. First consider two-step models. These assume that the listener initially computes some form of basic meaning, and then enriches this meaning to derive an implicature. Two-step models therefore predict that there is an extra processing step required in deriving the upper-bound relative to the lower-bound meaning. The extra step causes the processing cost observed by Bott and Noveck (2004) and others. One-step models, however, clearly cannot explain the cost to deriving the implicature by appealing to a different number of steps across interpretations. Nonetheless, there are general complexity differences across sentence forms that would suggest delays even without an extra step in the processing. For example, upper-bound sentences require dividing the subject set into a reference and a complement set (e.g., elephants that are mammals versus those that are not mammals), whereas lower-bound sentences only require a reference set (see Grodner, Gibson, & Watson, 2005, for evidence of additional processing cost when a reader must instantiate a complement set in addition to a reference set). Also, upper-bound sentences involve negation whereas lower-bound sentences do not, and processing negation may cause difficulty (see Clark & Chase, 1972). Finally, upper-bound sentences contain more information than lower-bound sentences (upper-bound sentences logically entail lower-bound sentences, but not vice versa), and their interpretation might therefore require more computation. According to one-step models, composition and comprehension processes are more complex for the upper-bound meaning, and processing the upper-bound sentence is therefore more time-consuming.

While both types of models predict delayed scalar implicatures, they make different claims about a participant's reasoning prior to the completed response. Two-step models propose that participants pass through a step in which they first access the lower-bound interpretation before enriching it to form the upper-bound, but one-step models maintain that participants will directly incorporate the upper-bound interpretation into the sentence representation. This paper presents the results of three sentence verification tasks that used a mouse-tracking paradigm to test between these models. The mouse trajectories provide vital information about participants' reasoning prior to their final response in a way that would not be observable with reaction times alone. As we explain in Experiment 1, the mouse path predictions are different for the one-step and two-step models.

Mouse tracking

Mouse-tracking has been used successfully to investigate a broad range of phenomena in domains as diverse as phonetics (Spivey, Grosjean, & Knoblich, 2005), syntactic processing (Farmer, Anderson, & Spivey, 2007), social cognition (Freeman & Ambady, 2010), and lie detection (Duran, Dale, & McNamara, 2010). See Freeman, Dale, and Farmer (2011), for a review. Participants' mouse movements during responses reveal that attentional processes and decision-making processes are not mutually exclusive and distinct; rather motor movements demonstrate online integration of both at their earliest steps of processing (Spivey, 2007).

In a typical mouse tracking experiment the cursor starts at the bottom of the screen in the center, with two response options placed in the top left and right corners—perhaps two pictures, or true and false response options (see Fig. 1). The participant hears or sees a sentence, then uses a mouse to move the cursor and click on the target response. The directness of the mouse trajectory from the

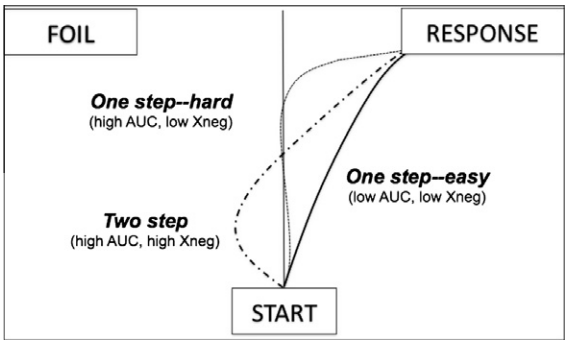

Fig. 1. Idealized mouse tracks for easy and hard one-step processes compared to a two-step process. The easy one-step trajectory (solid line) has a direct path towards the target response and does not enter into the competitor mouse space. This path will have low AUC (area under the curve) and low Xneg (deviation from the medial axis towards the competitor, FOIL). The hard one-step trajectory (light dotted line) pushes up the medial axis before veering towards the response (a so-called "T" motion). The AUC is correspondingly large, but because the mouse path barely crosses over into the competitor space, the Xneg is very small. Finally, the two-step path (dot-dashed) initially deviates towards the foil, but then returns to the response. The path is indirect and therefore has a high AUC, but because the path deviates into the competitor space it also has a high Xneg.

bottom of the screen to the target provides information about the underlying cognitive processes of the formulation and selection of a response. When participants are immediately confident of their response, mouse paths depart early from the central axis and progress directly to the target, as indicated by the idealized solid line trajectory in Fig. 1. Conversely, when participants have difficulty and/or are oscillating between two responses, they generally hover around or move the mouse up the center of the screen for a longer period resulting in a less direct mouse trajectory towards the target, as indicated by the idealized dotted line trajectory in Fig. 1 (e.g. Spivey et al., 2005). In contrast, if participants initially consider one option before ultimately selecting the other, mouse paths initially head away from the ultimate target, as indicated by the dash-dot trajectory in Fig. 1 (e.g., Dale & Duran, 2011). Thus the mouse path provides information about the interpretation processes that occur prior to the completed interpretation (the ultimate response).

The qualitatively different mouse tracks in Fig. 1 can be usefully characterized in terms of two quantitative measures, *area under the curve* (AUC) and *horizontal deviation towards the competitor response* (Xneg). AUC measures the geometric area between the actual trajectory and a straight line between the trajectory's start and end points. Mouse paths that deviate far from the shortest path will have a large AUC. AUC is thought to provide a general measure of processing difficulty (Freeman & Ambady, 2010), and the more difficult the task the greater AUC (Farmer et al., 2007; Freeman, Ambady, Rule, & Johnson, 2008; Spivey, 2007). Xneg measures deviation away from the medial axis towards the competitor response. A large Xneg indicates that the participant strayed far into competitor space. We report AUC and Xneg in our experiments. AUC allows our results to be compared with other mouse tracking experiments and Xneg provides a means of distinguishing between one-step and two-step processing models, as we describe below (further explanation of the difference between AUC and Xneg can be seen in Fig. 1).

Experiment 1

Experiment 1 compared lower-bound and upper-bound interpretations of sentences like *some elephants are mammals*. Participants read propositional statements modified by *some* or *all*, and decided whether each sentence was true or false. Participants classified six types of sentences, including the critical sentences, such as *some elephants are mammals*, and five types of control sentences. Critical sentences were true if responses were based on the lower-bound meaning and false if their responses were based on the upper-bound meaning. The five control sentence types were the same as those of Bott et al. (2012), and were designed so that participants could not predict whether the sentence was going to be true or false from the quantifier-subject relationship (see Table 1 for a complete list of sentence types). For example, on reading *some elephants are*..., participants were not able to predict the truth of the sentence before the final word, because some completions made the sentence true (e.g., *Indian*), while other completions made the sentence false (e.g., *insects*). Statements were presented one word at a time, and participants were free to move the mouse when the final word was presented. Responses were made by clicking on "T" (true) or "F" (false) targets at the top two corners of the screen.

We wanted to compare upper and lower-bound mouse paths for the critical sentences. We therefore introduced a training phase at the start of the experiment that biased participants into understanding the critical sentences according to one interpretation or the other (see Bott & Noveck, 2004, Experiment 1; and Rips, 1975). In the *logical* condition, participants received feedback encouraging a lower-bound, logical interpretation (a true response). In the *pragmatic* condition, feedback encouraged an upper-bound, pragmatic interpretation (a false response). Thus, participants in the logical condition received a message reading "correct" when they responded true to the *some* critical sentences, and a message reading "incorrect" when they responded false. Participants in the pragmatic condition received the reverse response mapping feedback. Participants also received feedback on the control sentences, which was the same for all participants. After the training phase, participants judged the truth of similar test sentences, but did not receive feedback.

Predictions for the models are as follows (see Fig. 1). Because previous studies have found lower-bound interpretations to be relatively fast, we predicted that participants in the logical condition would, on average, move the mouse directly to the target when responding to the critical sentences, as in the "one step – easy" mouse path in Fig. 1. In contrast, previous studies have found upper-bound interpretations to be relatively slow. According to one-step models, the delay in processing upper-bound sentences is caused by extra composition and comprehension processes, such as needing to form a

Table 1
Experimental stimuli.

	Name	Example	True/false
Experiments 1, 2 and 3	Some critical	Some elephants are mammals	Exp
	All true	All elephants are mammals	T
	All false	All elephants are insects	F
	Some true	Some mammals are elephants	T
	Some false	Some elephants are insects	F
Experiment 1 only	Some subordinate true	Some elephants are Indian	T
Experiment 2 only	All super false	All mammals are elephants	F
Experiment 3 only	No critical	No elephants are mammals	F
	No true	No elephants are insects	T
	No false	No elephants are Indian	F

Notes: All three experiments used sentences in rows 1–5. Individual experiments additionally used sentences from the appropriate rows. Exp refers to the experimental sentence, which could be either true or false.

complement set, process the negation, and verify the additional information. Participants in the pragmatic condition would therefore require more time to come to a decision about whether the sentence is true or false. In mouse tracking terms, they should therefore move their mouse further up the vertical axis before deviating towards the target response (as in the "one step – hard" mouse path in Fig. 1). According to one-step models, AUC for critical sentences should be larger for participants in the pragmatic condition than the logical condition, but Xneg should be near zero for both groups of participants. In contrast, according to two-step models, interpretations of the upper-bound sentence involve first accessing the lower-bound meaning before the completed implicature. Average mouse trajectories should therefore first move towards lower-bound targets (true) and then shift towards upper-bound targets (false) in a secondary step (as in the "two step" mouse paths in Fig. 1). According to two-step models, AUC may or may not be greater for upper-bound than lower-bound interpretations; however, if mouse trajectories are sensitive to the initial representations in a two-step process, then Xneg should be larger for upper-bound than lower-bound interpretations.

Method

Participants

Forty Cardiff University psychology students participated for course credit. All were native speakers of English.

Stimuli

Sentences frames were of the form "Quantifier X are Y". The quantifier was either *all* or *some*, and X and Y were either exemplars or categories, depending on the sentence type. Sentences were generated from 18 categories (e.g., mammals, insects, cars) and 30 exemplars (e.g., cows, bees, Ferraris). Experiment 1 used six different sentence types, as shown in Table 1. Appendix A shows a complete list of stimuli.

Design

Participants were presented with a biasing context that encouraged either an upper-bound or a lower-bound interpretation. The context consisted of a training phase in which feedback ("correct" or "incorrect") was given to reinforce either the upper-bound or lower-bound interpretations of *some*. Feedback was also given for the control sentences. Participants were randomly allocated to a pragmatic condition or a logical condition. No feedback was given in the experimental phase.

The training phase involved 100 sentences, 25 of which were critical items along with 15 *some true* items, 15 *some subordinate true* items, 15 *some false* items, 15 *all false* items, and 15 *all true* items. These items were different from the items used in the main experiment and were not rotated across conditions.

Each item for the test phase was formed by using an exemplar, e.g., "elephants", in one of six versions corresponding to the six sentence types above. Participants saw each exemplar in all six sentence types. In total, there were 30 exemplars by 6 conditions, making 180 sentences in the test phase. Sentences were presented in a different random order for each participant.

True and false response boxes (labeled "T" and "F") were presented at the top-left or top-right of the screen. Positioning of the response boxes was counterbalanced across participants but the same configuration was used throughout the experiment.

Procedure and equipment

The *MouseTracker* software (Freeman & Ambady, 2010) was used to run all of the experiments reported in this paper. We used standard monitors measuring 35 by 25 cm, and normal, laser mice with no mousepad. The start box was placed at the bottom center of the screen and measured 2×5 cm. The two response boxes we located at top left and top right corners of the screen at a distance of 1 cm from the edges. Both responses boxes measured 5×5 cm. The linear distance from the starting position of the cursor arrow in the middle of the start box to the edge of a response box measured 22 cm. The distance moved by the participant's hand to select a response was approximately 17 cm.

Participants were told to evaluate the truth of statements by clicking on the response boxes. They were also told to ask for clarification if they received feedback for a trial that did not make sense (although no participant did). No specific instructions were given about upper or

lower-bound interpretations of the scalar term (cf. Bott & Noveck, 2004).

Participants started the trial by clicking on the start box at the bottom center of the screen. Sentences were presented word by word in the middle of the screen at a rate of 300 ms per word. Mouse tracking began at the onset of the final word in the sentence. If participants had not initiated any mouse movement within 500 ms of the final word, they received a warning telling them to respond more quickly.

Results

Preprocessing

For all of the experiments presented in this study, responses were considered outliers if mouse trajectories were three standard deviations outside of the mean AUC of all responses. We also removed incorrect responses from all conditions. In Experiment 1, 1.5% of responses were excluded as outliers and 7% as incorrect responses. Less than 1% of responses had initiation times greater than 500 ms. These responses were also removed.

To compare the mouse movements of responses with different response times, for example a 900 ms response versus a 1200 ms response, we normalized the time course of the mouse paths into 101 time steps, as is standard in mouse tracking (Freeman & Ambady, 2010). The time steps also provide a rough space of potential areas of processing during the response: response initiation stages (time steps 1–25), early/middle stages (time steps 26–50), middle/late stages (time steps 51–75), and late stages (time steps 76–101). In general, the majority of response initiation stages cluster around the start button, that is participants rarely initiate their response in the first few time steps, and the majority of time steps in the late stages cluster around the response button because participants generally stop moving the mouse before initiating a click. The horizontal coordinates corresponding to each time quadrant and each condition are shown in Appendix B.

At each time step the position of the mouse was represented by (X,Y) coordinates, with X ranging from -1 to 1 and Y ranging from 0 to 1.5. This 2×1.5 rectangle roughly corresponds to classical CRT monitors and facilitates comparisons across different screen resolutions. The *Analyser* program in the MouseTracker suite (Freeman & Ambady, 2010) was used to perform the normalization.

Mouse tracking analysis

Figs. 2 and 3 display average mouse paths for participants in the logical and pragmatic conditions respectively, collapsed across the counterbalanced right and left positions of the targets. Mouse paths for the critical sentences are shown by the dark crosses. In the logical condition the critical sentences were true, and the mouse paths display a simple arc towards the true target. Conversely, the critical sentences for the pragmatic condition were false. Here, however, the mouse paths deviate substantially towards the true target before crossing back over the medial axis towards false. While several types of control sentences deviate slightly towards the competitor before returning to the correct target, such as the *some subordinate true* sentences in Fig. 3, these deviations appear much smaller and occur earlier in the paths than the deviation for the upper-bound sentences. In general, all control trajectories shift towards false in the pragmatic compared to the logical condition, consistent with the shift in overall proportions of false versus true responses in the two conditions.

Mean accuracy, AUC, and reaction times are shown in Table 2. Mean AUC scores for true control sentences were all higher in the pragmatic condition than the logical condition, while false sentences exhibited the opposite pattern. AUC scores for critical sentences were higher in the pragmatic condition than in the logical condition. The effect of training condition on AUC scores for critical sentences was assessed by fitting a linear mixed-effects regression model (Baayen, Davidson, & Bates, 2008) with sentence type (*some critical*, *some false*, *some subordinate true*, *some true*, *all false*, *all true*) and training (*pragmatic versus logical*) as categorical fixed effects, along with the interaction between these two variables. The variables were dummy coded so that *some critical* sentences in the pragmatic condition functioned as a baseline, and the "main effect" of training measured the difference between *some critical* sentences in the pragmatic and logical conditions. Random effects were included for subjects and items. A Markov Chain Monte Carlo (MCMC) method was used to obtain *p*-values for fixed effects as recommended by Baayen et al. AUC scores for critical sentences were significantly greater for the upper-bound interpretations of the pragmatic condition than for the lower-bound interpretations of the logical condition, $t = 17.79$, $p < .001$, suggesting greater overall difficulty for upper-bound interpretations.

We also analyzed the *x*-coordinates to establish whether participants deviated into the competitor half of the trajectory, that is, whether there was a significant Xneg. For each condition, we calculated the mean deviation into competitor space at each time point. Next, we found the time point with the maximum deviation and tested whether it was significantly different from zero, using both participant and item variability. For pragmatic participants, the largest mean deviation for the critical sentences, i.e., the largest Xneg, occurred at time step X_43, and had *x*-coordinate $M = -.24$. A one sample *t*-test showed that this was significantly different from zero for participants, $t_1(19) = 6.03$, $p < .001$, and items, $t_2(29) = 9.98$, $p < .001$. In contrast, there were no Xneg scores for the critical sentences in the logical condition (i.e., the mean trajectory for the logical sentences did not cross into the competitor half of the space).

Several of the control sentences also displayed deviations into the competitor space. Specifically, *all false* and *some false* in the logical condition and *some subordinate true* in the pragmatic condition. These deviations were smaller than those associated with the critical sentences, all $t_1(19) < 3.06$, p's $< .001$, all $t_2(29)$'s > 3.44, p's $< .001$, and they occurred much earlier in the trajectory: the control deviations occurred at X_23, X_25, and at X_28, whereas the critical deviations occurred at X_43. These effects are discussed below.

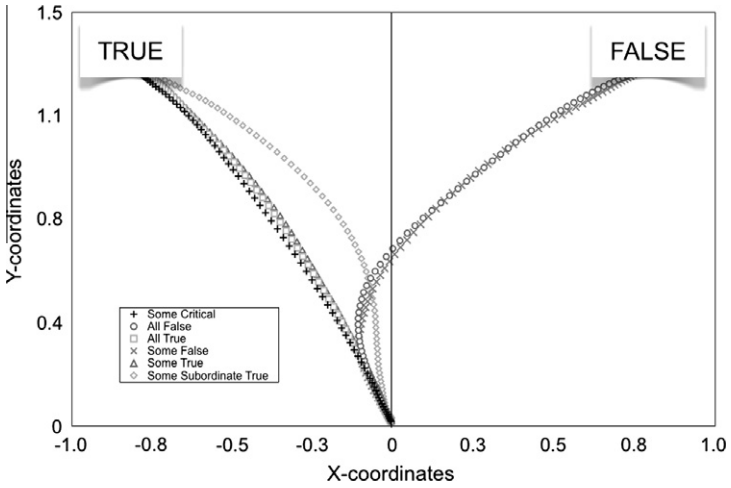

Fig. 2. Average mouse trajectories in Experiment 1 for the logical condition. Data points represent the average X and Y positions at each of 101 time steps.

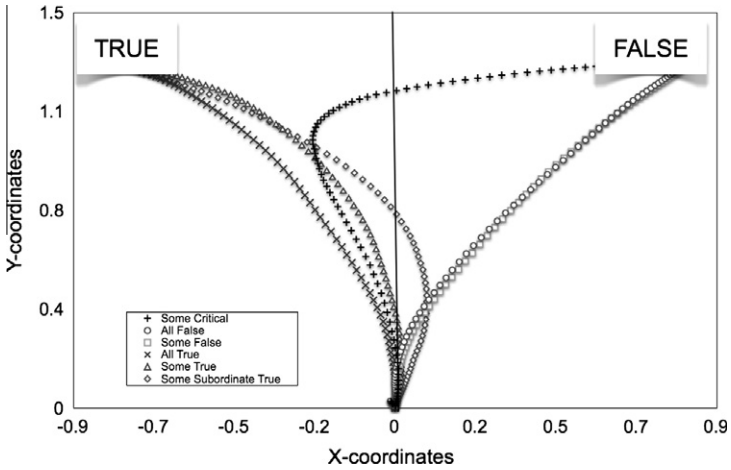

Fig. 3. Average mouse trajectories in Experiment 1 for the pragmatic condition. Data points represent the average X and Y positions at each of 101 time steps.

Table 2
Experiment 1: Mean area under the curve (AUC), accuracy, and reaction time (RT) for logical and pragmatic conditions.

Sentence type[a]	Logical			Pragmatic		
	AUC *M*(*SD*)	Accuracy (%)	RT *M*(*SD*)	AUC *M*(*SD*)	Accuracy (%)	RT *M*(*SD*)
All true	0.30(.84)	97	860(312)	0.71(1.25)	96	1035 404)
All false	0.84(1.22)	98	979(350)	0.37(1.22)	98	949(322)
Some true	0.34(0.92)	98	876(322)	0.91(1.41)	94	1006(442)
Some false	0.84(1.18)	97	975(326)	0.32(0.83)	98	947(307)
Some subordinate true	1.16(1.25)	95	1007(363)	1.44(1.61)	96	1092(390)
Some critical[b]	0.25(0.8)	97	856(320)	2.26(1.48)	87	1295(426)

[a] $n = 30$ items per sentence type.
[b] True in logical condition, false in pragmatic condition.

Discussion

Mouse paths for the upper-bound interpretations of critical sentences deviated towards the competitor response (true) prior to the correct response (false). No such deviation, however, was observed for lower-bound interpretations. Initially, average mouse trajectories for upper-bound responses were reliably heading towards true and only later corrected towards false. Because participants initially moved their mouse movements towards true, this strongly suggests that a lower-bound meaning was active early, and the upper-bound meaning was activated

relatively late in processing. These results are consistent with two-step models, in which participants begin with a lower-bound interpretation and only subsequently derive an upper-bound interpretation for the critical sentences. The results are not consistent with one-step models, which predict monotonic trajectories in the direction of false responses for upper-bound interpretations.

We also observed some small and early deviations into the competitor space for some of the control sentences. Our view is that these were due to response biases caused by an imbalance in the ratio of true to false responding in the training phase. In the logical condition, there was a predominance of true sentences overall, hence participants displayed negative mouse paths whenever they responded false. In the pragmatic condition, while there was a more equal ratio of true to false sentences overall, sentences that involved *some* and a subordinate item (e.g., *elephants*) were predominantly false, hence there were initial trajectories towards false. Furthermore, to foreshadow the results of Experiment 2, we do not replicate the negative trajectories when we reduced the imbalance in true/false responding.

In this study, participants were trained to respond to the critical sentences in a particular way. It is conceivable that participants might ordinarily be inclined to respond true to our experimental sentences, but in the pragmatic condition they responded false in order to conform to the experimental demands introduced by the feedback during the training phase. This could explain an initial deviation towards the true response, followed by a later, metalinguistic deviation towards false. In order to test the possibility that our results were an artifact of task demands, we repeated Experiment 1 but without training participants to respond a particular way to critical sentences.

Experiment 2

Experiment 2 compared lower-bound and upper-bound interpretations of sentences in an unbiased context. Participants had to verify categorical sentences, just as they did in Experiment 1, but they did not receive feedback on their responses to the critical sentences. There were therefore no experimenter demands to make false responses to the critical sentences and participants could respond with their most natural interpretation. Participants underwent a practice phase, just as they did in Experiment 1, but critical sentences were not presented.

We introduced one further change in this experiment relative to Experiment 1. In Experiment 1, there were three true control sentences and two false control sentences. This meant that there was a response bias towards true responding, especially in the logical condition after training. While it is unlikely that response biases could explain the trajectories that we observed in Experiment 1 for critical sentences (the effects on critical sentences were later and stronger than those apparently associated with response biases for control sentences), we wanted to eliminate this possibility. We therefore replaced one of the true control sentences, *some subordinate true*, with another false control condition, *all super false*, such as *all mammals are elephants* (this meant that the control sentences we used were identical in form to those of Bott & Noveck, 2004). See Table 1 for a complete list of sentence types. We also adjusted the proportions of the sentences assigned to each condition so that there was an equal ratio of true to false items overall (100 trials in the practice phase, 50:50 true:false). The ratio of true to false was also balanced within each quantifier so that there was an approximately equal ratio within *some* sentences (25:20) and within *all* sentences (25:30). The critical sentences were not included in these calculations because participants were free to choose which interpretation they preferred.

Method

Participants

Thirty-two undergraduate students from Cardiff University participated in Experiment 2 in exchange for cash payment. All were native speakers of English.

Stimuli

We used the same stimuli base as Experiment 1, but changed one of the six sentence types. See Table 1 for a list of sentence types and examples.

Design

All participants underwent a practice phase of 100 trials. These trials consisted of 15 *all super false* sentences, 25 *some true* sentences, 20 *some false* sentences, 15 *all false* sentences, and 25 *all true* sentences. The breakdown was done in this way to ensure that each quantifier had roughly an equal chance of being true or false. No critical sentences were included in the practice session.

The main experiment contained 180 trials. There were six conditions and all conditions had an equal number of trials. Thirty exemplars were selected equally across different category types and participants were presented with the six different sentence types built around each exemplar (see Appendix A). Participants did not receive any feedback during the main experiment.

Procedure

The experimental procedure was the same as Experiment 1, although participants were not trained on critical items.

Results

Preprocessing

We removed 1.8% of the responses as outliers. A further 3% incorrect responses were removed from the control conditions. Less than 1% of the responses had initiation times greater than 500 ms.

Accuracy

Accuracy rates are shown in Table 3. Participants answered the control questions extremely accurately, as they did in Experiment 1. For the critical sentences, participants responded with a slight bias towards upper-bound

Table 3
Experiment 2: Mean area under the curve (AUC), accuracy, and reaction time (RT).

Sentence type[a]	AUC *M*(*SD*)	Accuracy (%)	RT *M*(*SD*)
All true	0.84(1.47)	93	1055(508)
All false	0.29(.79)	99	946(333)
All super false	1.09(1.55)	92	1141(484)
Some true	0.80(1.44)	95	1131(493)
Some false	0.33(.91)	98	975(368)
Some critical (true responses)	0.84(1.45)	41	1060(518)
(false responses)	1.89(1.78)	59	1260(533)

[a] $n = 30$ items per sentence type.

interpretations (false) but the sample still generated a high proportion of lower-bound responses. Overall the accuracy rates are similar to Bott and Noveck (2004; Experiment 3), who also observed a 59% upper-bound response rate.

Mouse tracking analysis

Fig. 4 shows the average mouse trajectories for the false and true responses respectively, collapsed across the counterbalanced right and left positions of the targets. For the critical sentences, the data are the average mouse paths of all the responses to false (upper-bound interpretation) or true (lower-bound interpretation), independently of which participants made which responses. For the control sentences, the data show the average of the correct responses only. The pattern is very similar to the results of Experiment 1: Upper-bound responses deviate towards the competitor response, whereas lower-bound responses do not.

Mean accuracy, AUC, and reaction times are shown in Table 3. AUC scores varied across the various sentence types, and on critical sentences the AUC scores were higher for false (upper-bound) responses than true (lower-bound). We analyzed AUC scores using a repeated measures, mixed model design. Categorical fixed effects were quantifier (*all*, *some control*, *some critical*), and response (*true*, *false*), along with the interaction between these two variables. Sentence type was entered as a single categorical fixed effect with the seven levels shown in Fig. 4, including separate levels for true and false responses to some critical sentences. Sentence type was dummy coded with pragmatic responses to *some critical* sentences serving as a baseline. Random effects were included for subjects and items. AUC was significantly greater for upper-bound than lower-bound interpretations, $t = 11.40$, $p < .001$, again suggesting greater deviation for upper-bound interpretations.

In the analysis of Xneg, for the upper-bound (false) responses, the largest negative deviation was at time step X_41, $M = -.27$, which was significantly different to zero, $t_1(28) = 3.78$, $p < .001$, $t_2(29) = 6.99$, $p < .001$. The data from 3 participants were not included in this analysis because they responded true to all of the critical sentences. There were no deviations into the competitor space for the lower-bound (true) responses.

One potential concern with the repeated measures analysis is that it includes participants who might be very unsure of their responses. Participants who are unsure may deviate towards either of the response options prior to the final decision. This effect may be much greater for upper-bound responses (false) because participants could have a general response bias towards true. To eliminate this explanation of our findings, we divided participants into logical responders and pragmatic responders depending on whether the majority of their responses (65%) were

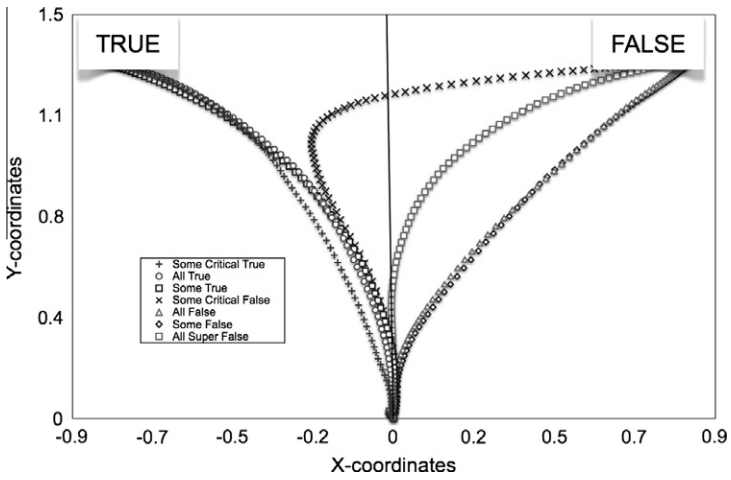

Fig. 4. Average mouse trajectories in Experiment 2. Data points represent the average X and Y positions at each of 101 time steps. Diagonal crosses to false correspond to upper-bound interpretations of the critical sentences and vertical crosses to true correspond to lower-bound interpretations.

lower-bound (logical) or upper-bound (pragmatic). This generated two groups of participants who were relatively confident about their responses and on which we could conduct an analysis similar to that of Experiment 1.

Classifying on the basis of the majority of responses resulted in 14 out of 32 logical responders, 14 pragmatic responders, and four participants who had equal numbers of upper and lower-bound responses. This latter group were removed from the analysis because they could not be classified. We also removed the minority "incorrect" responses from the logical and pragmatic groups. Average mouse trajectories for pragmatic responders in the false response conditions are shown in Fig. 5. As before, AUCs for false responses to critical items differed significantly from true responses to critical items, $t = 12.10$, $p < .001$. In the analysis of Xneg, the pragmatic responders significantly deviated towards true at X_41, $M = -.19$, $t_1(13) = 4.22$, $p < .001$, $t_2(29) = 6.77$, $p < .001$. The logical responders did not deviate towards false when responding true, and none of the control conditions from either group had any negative coordinates. In short, the between-subject analysis leads to the same conclusion as the within-subject analysis.

Discussion

Experiment 2 examined average mouse trajectories for spontaneous lower-bound and upper-bound interpretations of critical sentences. The mouse paths for upper-bound responses were similar to those of Experiment 1, in which participants were trained to interpret these sentences one way or another. In both experiments mouse paths of the upper-bound responses deviated towards the competitor response before completing their trajectories towards false, but the same pattern was not observed for the lower-bound responses. These results rule out the possibility that the nonmonotonic mouse paths seen in Experiment 1 were due to an artifact of the training procedure. Experiments 1 and 2 are therefore consistent with participants accessing lower-bound meanings before upper-bound meanings, as predicted by two-step models.

One explanation for our findings is that the initial attraction towards the true target might arise from a mismatch between the truth of the embedded proposition (*elephants are mammals*, true) and the truth of the quantified sentence (*some elephants are mammals*, false in the pragmatic condition). If participants delayed processing of the quantifier until after they had fully processed the embedded proposition, they could display initial deviations towards true in exactly the pattern that we observed. Delayed processing of the quantifier would be a sensible processing strategy within our task and in other, more naturalistic situations. Content words carry by far the most information in any conversational exchange and it makes sense for listeners to focus their attention on these components of the sentence. Furthermore, quantifiers need elements over which they quantify, and to some degree, people cannot fully interpret quantifiers until they know what it is that is being quantified. Although we did not observe delayed quantifier effects on any of our control sentences (see Bott & Noveck, 2004, for a similar point), it is conceivable that the propositions within these sentences were confounded by idiosyncratic factors that differentiate them from the critical sentences. In particular, the subject-predicate relationship was different in these control sentences (supercategory-exemplar) than it was in the *some* critical sentences (exemplar-supercategory). We therefore designed Experiment 3 to test whether nonmonotonic mouse paths would be observed with the sentences that involved the same form of embedded proposition as the critical sentences but used a different quantifier.

Experiment 3

In Experiment 3 we compared the critical sentences from Experiments 1 and 2 with sentences that used *no* as quantifier but involved the same embedded proposition, such as *no elephants are mammals*. These latter sentences

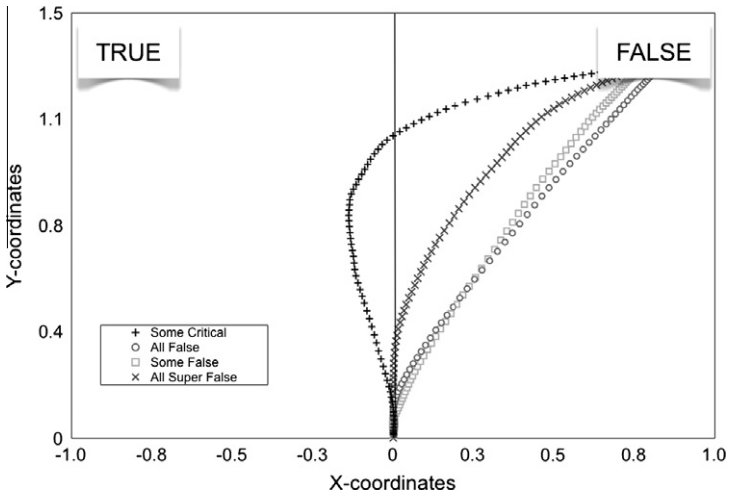

Fig. 5. Average mouse trajectories in Experiment 2 of the pragmatic responders for the false conditions. Data points represent the average X and Y positions at each of 101 time steps.

constituted the *no critical* condition (see Table 1 for a complete list of sentence types). All participants received upper-bound training, just as in the pragmatic condition of Experiment 1, so that the *some* critical sentences were false. The *no* quantifier reverses the truth of the embedded sentence, just as with the upper-bound interpretation of the *some* critical sentences, but it does not involve a scalar implicature. If the nonmonotonic mouse paths are due to the mismatch between sentence truth and embedded proposition, nonmonotonic mouse paths should also be observed with the comparable *no* sentences.

Method

Participants

Twenty-four undergraduate students from Cardiff University participated in exchange for cash payment. All were native speakers of English.

Stimuli

Eight sentence types were used in Experiment 3 (see Table 1).

Design

The experiment was a within participant design. All participants underwent a practice phrase of 100 items, which consisted of 10 *all false* items along with 12 *some true* items, 10 *some false* items, 20 *some critical* items, 12 *all true* items, 12 *no critical* items, 12 *no false* items, and 12 *no true items*.

The main experiment contained 206 trials. Of these, 176 were experimental trials and the remaining 30 were filler trials. The experimental trials involved counter-balanced items distributed equally across the eight conditions (22 trials per condition). Filler trials were included to counter-balance the high proportion of false responses in the rest of the design. The filler trials were all of the *all true* form but the items were novel and not rotated across the conditions of the experiment.

Procedure

The experimental procedure was the same as the Experiment 1 pragmatic condition in that participants were trained to respond to *some* critical items as *false*.

Results

Preprocessing

We removed 1.3% of responses as outliers. A further 6% of responses were removed because they were incorrect. Table 4 shows the accuracy rates for all conditions. Less than 1% of the responses had initiation times greater than 500 ms.

Mouse tracking analysis

Figs. 6 and 7 show the average mouse trajectories for the false response conditions and true response conditions, respectively, collapsed across the counterbalanced right and left positions of the targets. The pattern for the *some critical* sentences is very similar to that in Experiment 1: Upper-bound responses deviate towards the true responses before arriving at false responses. Of principle interest for this experiment, however, were the *no critical* sentences. The average trajectory for these sentences does not look like the trajectory for the *some critical* sentences. Only the *some critical* sentences show deviation towards the competitor response; the *no critical* sentences push up the vertical axis initially, but when the trajectory deviates from the axis it heads for the target response, not the competitor response.

Table 4
Experiment 3: Mean area under the curve (AUC), accuracy, and reaction time (RT).

Sentence type[a]	AUC *M*(*SD*)	Accuracy (%)	RT *M*(*SD*)
All true	0.67(1.39)	94	1070(390)
All false	1.28(1.47)	92	930(380)
Some true	1.32(1.40)	90	1034(390)
Some false	0.57(1.13)	97	943(328)
Some critical (false)	2.08(1.88)	78	1188(423)
No critical (false)	1.02(1.46)	93	1154(508)
No true	2.34(1.60)	93	1241(437)
No false	1.47(1.32)	95	1150(417)

[a] $n = 22$ items per sentence type, plus 30 *all true* filler trials (not analyzed).

The AUC for each of the conditions is shown in Table 4. We analyzed AUC using a categorical fixed effect of sentence type with the eight levels shown in Figs. 6 and 7, dummy coded with *some critical* sentences serving as the baseline reference category. Random effects were included for subjects and items. AUC was significantly greater for *some critical* than *no critical* sentences, $t = 11.12$, $p < .001$.

We further analyzed the *x*-coordinates to test for deviation into the competitor half of the trajectory. For the *some* critical sentences, the largest mean deviation occurred at time step X_40, $M = -.23$. A one sample *t*-test showed that this was significant for participants, $t_1(23) = 5.47$, $p < .001$, and items, $t_2(21) = 10.21$, $p < .001$. Conversely, the *no critical* sentences did not show any deviation into the competitor space.

Two control conditions showed negative mean *x*-coordinates during part of the average response paths. First, the *all false* condition, shown in Fig. 6, had a slight deviation towards true, the largest average deviation being at time step X_30, $M = -.07$. However, one-sample *t*-tests revealed the pattern was not significant, $t_1(23) = 1.65$, $p = .11$, $t_2(21) = 1.75$, $p = .09$. The trend could reflect the result of having 30 extra *all true* items as fillers to counterbalance an otherwise overwhelming bias towards false responses across sentence types. Second, the *no true* conditions, such as *no elephants are reptiles*, also showed deviation away from the target. The largest average deviation was at time step X_46, $M = -.30$ and the one sample *t*-test shows that this was significant for both subjects, $t_1(23) = 7.42$, $p < .001$, and items, $t_2(21) = 17.49$, $p < .001$. We discuss this result below.

Discussion

Experiment 3 tested whether the nonmonotonic mouse paths in the *some* critical sentences were because the truth

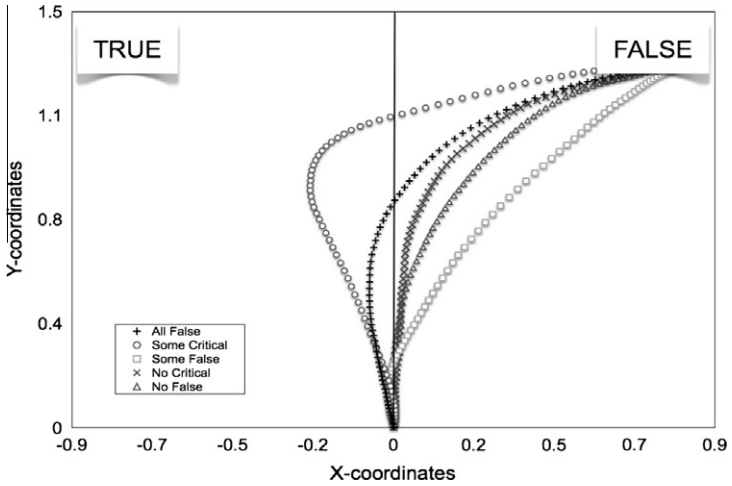

Fig. 6. Average mouse trajectories in Experiment 3 for the false response conditions. Data points represent the average X and Y positions at each of 101 time steps.

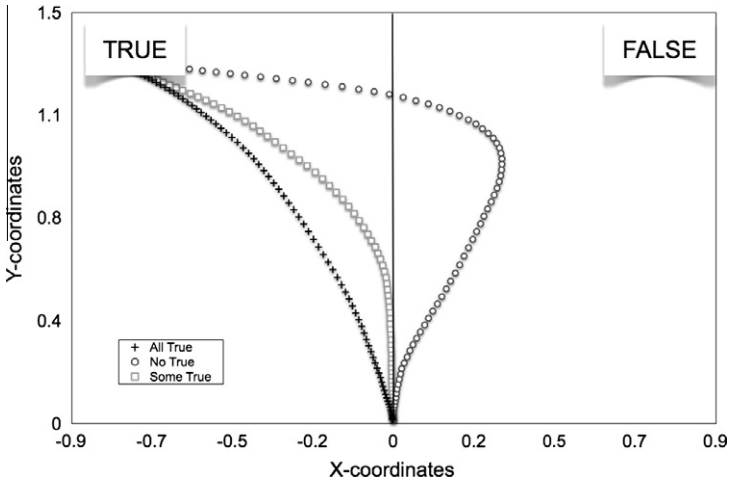

Fig. 7. Average mouse trajectories in Experiment 3 for the true response conditions. Data points represent the average X and Y positions at each of 101 time steps.

of the embedded proposition was different to the truth of the quantified sentence. We compared *some* critical sentences against other sentences that had the same incongruency between sentence truth and embedded proposition and the same exemplar–category relationship, but lacked the scalar implicature. These were sentences that involved *no*, as in *no elephants are mammals*. Whilst we replicated the nonmonotonic mouse paths for the *some* critical sentences, we did not find evidence of nonmonotonic paths for comparable *no* sentences, which would be expected if the effects we observed with *some* were due to an interference between sentence truth and embedded proposition.

Interestingly however, the *no true* sentences, such as *no elephants are reptiles*, showed significant deviations away from the target. Participants initially directed their mouse towards false, and then reversed the trajectory to click on true. Why was there a difference in mouse paths between the *no true* and the *no critical* sentences? We can think of several potential explanations. First, it is conceivable that initial deviations towards false were the result of a response bias for *no* sentences towards false (as with *some* sentences, twice as many *no* sentences were false compared to the number that were true). However, one would then also expect a response bias effect on *some true* sentences, but that is not what the data show. Below we suggest an alternative explanation based on the quantifier being partially delayed.

We first consider two extreme versions of how participants might process quantifiers in our task. As discussed at the end of Experiment 2, it is possible that participants might initially bypass the quantifier completely, storing it in a buffer, and directly process the embedded proposition. After processing the proposition, they could return to derive the complete, quantified, sentence meaning. Given the results of the *no critical* sentences, however, this seems

unlikely: if participants were evaluating the proposition first, we would have observed initial mouse paths towards true for these sentences. At the other extreme, participants, might process the proposition incrementally, starting with quantifier and then proceeding with the rest of the proposition, consistent with standard psycholinguistic theories about incremental language processing (e.g., Just & Carpenter, 1980). However, the mouse paths for the *no true* sentences argue against this. The initial deviation towards false suggests that, if anything, participants first analyzed the embedded proposition, resulting in a false trajectory, and only later did participants integrate the quantifier and reverse the trajectory.

Between the previous two extreme processing accounts however, the possibility exists that the quantifier was *partially* processed prior to verifying the embedded proposition (see Urbach & Kutas, 2010, for a similar conclusion about quantifier processing). In the *no* conditions, participants could have stored the meaning of the quantifier *no*, then interpreted the proposition, before finally integrating the two after both had been processed. In the *no critical* condition, the negation and the true proposition would have been integrated without any difficulty, but in the *no true* condition, combining negation with the false proposition could have been problematic because of the "double negative" (see, e.g., Clark & Chase, 1972). In the *no critical* condition, participants were able to quickly establish the sentence meaning and so did not deviate from the ideal trajectory. Because of the delay in integrating the quantifier and embedded proposition in the *no true* condition, however, participants initially directed their mouse towards the result of evaluating the embedded proposition during early stages of processing. This explanation is clearly post hoc and would require further experimentation to establish whether it is correct.

Could partial processing of the quantifier explain the nonmonotonic mouse paths observed for upper-bound sentences? Even if quantifier processing was partially delayed, the results of the *no critical* condition reveal that the quantifier was not processed so late that true exemplar–category propositions exerted an early influence on the mouse paths. Since the embedded proposition in the *no critical* sentences was identical to that of the *some critical* sentences, and there was no double negation in the *some critical* sentences, it is implausible that the nonmonotonic mouse paths in the *some* critical sentences were due to excessively late processing of the quantifier.

General discussion

Our goal in these experiments was to test between one- and two-step processing models of scalar implicatures. According to one-step models, the appropriate interpretation is incorporated directly into the sentence representation. Differences in processing times between upper and lower-bound interpretations are assumed to result from extra complexity in upper-bound interpretations relative to lower-bound interpretations. In contrast, two-step models attribute differences in processing times to an extra processing step inherent in the derivation of implicatures. According to the two-step model, lower-bound literal interpretations are fundamental and are generated automatically. When needed, upper-bound interpretations are subsequently derived from the corresponding lower-bound interpretations. Across three experiments, we found that when participants made upper-bound interpretations their mouse movements first deviated towards the lower-bound response option before targeting the upper-bound response option. In contrast, when participants made lower-bound interpretations their mouse movements went directly towards the target. In short, the results support a two-step model of implicature processing.

Implications for models

Our results suggest that participants interpret upper-bound meanings in two steps but lower-bound meanings in a single step. There are a variety of interpretations for what the steps might mean cognitively, however. We consider the results from the perspective of a classical Gricean account, Relevance (Sperber & Wilson, 1986), grammatical models of scalar implicatures as typified by Chierchia (2004, 2006, Chierchia, Fox, & Spector, 2008), and constraint-based models of scalar implicatures.

Gricean accounts

Our data are generally consistent with a direct implementation of the classical Gricean account of conversational implicature (1975). According to this view, when people derive upper-bound interpretations, they first decode the words and apply the normal rules of semantic composition to derive a coherent, literal sentence interpretation (Step 1). Next, appropriate components of the Cooperative Principle are considered, and the processor concludes that relevant alternatives to what the speaker said are presumed to be false (Step 2). Applying this theory to our data, initial trajectories towards the true target for sentences such as *some elephants are mammals* correspond to an evaluation of the literal meaning of these sentences (Step 1), and the redirection towards false corresponds to the application of the Cooperative Principle and general reasoning, when the listener concludes in Step 2 that the implied claim *some* [but not all] *elephants are mammals* is false.

Although a simple Gricean account is consistent with these data, it is not clear that it is sufficient as a cognitive processing model. The classical Gricean account assumes strict modularity and sequential processing (Step 2 cannot occur without the output of Step 1). This is somewhat counter to current conceptions of psycholinguistics that emphasize the incrementality of comprehension. Our data are certainly not compatible with a pure modular view that requires decisions to be complete before response motor action is initiated, because early deviation towards the competitor target indicates initiation of motor action before a final response decision (a similar point is made by Spivey (2007)). A less modular, more incremental possibility is that Step 2 processing could begin before the output of Step 1 was completed. In that case, the use of pragmatic principles could be integrated into the process at a much

earlier stage than a literal Gricean account would suggest. Future research needs to examine more precisely the point at which pragmatic enrichment begins.

Relevance

According to Relevance theory (Sperber & Wilson, 1986), pragmatic understanding involves a process of representational enrichment following an initial decoding of the message's conceptual content. The enrichment process is assumed to proceed incrementally until some criterion level of relevance is reached. From this perspective, scalar implicatures involve an initial decoding stage in which context-independent (logical) meaning is retrieved, followed by an enrichment process that derives implicature (Carston, 1998; Noveck & Sperber, 2007).

Relevance could predict our data if the output of the enrichment process was slow to come online. Initial trajectories towards true might correspond to a stage in which the decoding process had been completed, but that the enrichment process was still under way. Initial decoding suggested a true judgment and it was only later, after the enrichment process had reached some threshold level, that mouse trajectories reversed and moved towards false. Of course, this account requires a specification of why the enrichment is delayed in its output. A Gricean account might assume that participants were computing alternatives, comparing the informativeness of those alternatives, and engaging in deductive inference – all of which could be seen as time consuming processes – but these are not the sorts of processes that are suggested by Relevance, and it is not clear what the alternatives are.

Compositional accounts

An alternative approach to Relevance or the Gricean model is to consider scalar implicatures to be part of compositional semantics, rather than post-compositional pragmatics (see Chierchia, 2004, 2006; Chierchia et al., 2008). According to this view an implicit *only* operator applies to the scalar term at the earliest possible opportunity, that is, during the compositional process. Such accounts propose that the *only* operator activates some precompiled set of operations to exclude stronger alternatives. Our results suggest that the *only* operator must be applied at a fairly late stage in the compositional process, so that an initial interpretation of the sentence is formed before the *only* operator is applied, driving initial mouse trajectories towards the true target before the *only* operator kicks in. Crucially, because the *only* operator is part of the composition process, this account requires emerging interpretations to be accessible throughout the composition process, or at least before composition is complete.

Constraint-based models

The one-step model described in the Introduction was inspired by constraint-based and statistical models of language (e.g., Bates & MacWhinney, 1989; MacDonald et al., 1994; Traxler, Pickering & Clifton, 1998; van Gompel et al., 2005). Although there are no detailed constraint-based theories of scalar-implicature in the literature, we considered it plausible that a model that uses only statistical and frequency-based information might be able to predict when the *not all* inference arises and integrate this into the sentence representation. Crucially, such a model may not need standard Gricean pragmatic processes, such as computations of entailment relations, informativeness, and deduction. Nor would there be a need to represent maxims or rules of communication in the model – maxim-based behavior would be an emergent property of such a network. In this sense a constraint-based pragmatics model would be similar to connectionist models of English past tense that reproduce regularities in verb endings without explicit rules (e.g., Rumelhart & McClelland, 1986), or models that reproduce grammatical structure without an explicit representation of syntax (e.g., Elman, 1991). In our view the most straightforward prediction from such a model would be that the *not all* component of *some* would be incorporated directly into the sentence representation in a single step, and our experiments conflict with this prediction and therefore this type of model.

Nonetheless, the one-step model may characterize only a very simple implementation of a constraint-based architecture. More complex constraint-based models, such as those seen in the syntactic processing literature, might be able to explain our data. One of the properties of constraint-based models that might prove important is their ability to consider multiple interpretations simultaneously. Moreover, constraint-based models of syntactic parsing often assume that multiple syntactic interpretations are kept active throughout sentence comprehension, even when some of them are unlikely (e.g., Elman, Hare, & McRae, 2004; Farmer et al., 2007; MacDonald et al., 1994; McRae, Spivey-Knowlton, & Tanenhaus, 1998; Tabor & Tanenhaus, 1999). For example, consider the garden path sentence, ''the waiter served a steak enjoyed it immensely''. The sentence is temporarily ambiguous up until ''enjoyed'' because ''served'' could be treated as a main verb (the waiter was serving) or as a past participle (the waiter was served). According to Farmer et al., when a sentence like this is parsed, both interpretations would be kept active throughout sentence comprehension even though activation of the main verb interpretation would be very low after ''enjoyed''. Farmer et al. showed that individual-trial mouse trajectories responding to such sentences reflected the influence of both parsing options simultaneously (also see Spivey et al., 2005).

Multiple active interpretations could potentially provide a kind of one-step account for the nonmonotonic mouse paths we observed, but only on the assumption that different sources of relevant information become available at different times (e.g., Elman et al., 2004; McRae et al., 1998). A constraint-based model in which information favoring the upper-bound interpretation was delayed relative to information supporting the lower-bound interpretation could produce mouse trajectories that first headed towards the lower-bound response before changing direction as the relevant upper-bound information came online (rather like the Relevance model we suggest above). The delayed information might be consistent with elements of a Gricean model, such as the epistemic status of the

Table A2.1
Experiment 1: Mean *X*-coordinates for logical and pragmatic conditions across normalized time bins coordinates.

Condition and sentence type	Bin 1 (1–25) *M*(*SD*)	Bin 2 (25–50) *M*(*SD*)	Bin 3 (50–75) *M*(*SD*)	Bin 4 (75–101) *M*(*SD*)
Logical				
All true	−0.07(.17)	−0.39(.22)	−0.72(.22)	−0.82(.09)
All false	−0.06(.16)	0.05(.19)	0.60(.35)	0.83(.08)
Some true	−0.05(.14)	−0.34(.26)	−0.71(.31)	−0.84(.08)
Some super true	−0.04(.14)	−0.14(.28)	−0.57(.33)	−.084(.08)
Some false	−0.06(.16)	0.18(.30)	0.62(.33)	0.83(.08)
Some critical	−0.07(.14)	−0.39(.24)	−0.73(.34)	−0.83(.08)
Pragmatic				
All true	−0.06(.14)	−0.23(.26)	−0.61(.34)	−0.83(.1)
All false	0.03(.11)	0.30(.21)	0.66(.29)	0.82(.09)
Some super true	0.04(.14)	0.02(.37)	−0.48(.36)	−.81(.16)
Some true	0.05(.13)	−0.24(.21)	−0.50(.41)	−0.81(.1)
Some false	−0.03(.12)	0.32(.23)	0.67(.26)	0.83(.08)
Some Critical	−0.04(.18)	−0.22(.32)	0.16(.43)	0.79(.13)

Note: Negative values indicate movement towards the true target, and positive values indicate movement towards the false target.

Table A2.2
Experiment 2: Mean *X*-coordinates across normalized time bins coordinates.

Conditions	Bin 1 (1–25) *M*(*SD*)	Bin 2 (26–50) *M*(*SD*)	Bin 3 (51–75) *M*(*SD*)	Bin 4 (76–101) *M*(*SD*)
All true	−0.01(.17)	−0.18(.34)	−0.55(.43)	−0.80(.13)
All false	0.001(.13)	0.25(.32)	0.69(.34)	0.83(.12)
All super false	0.01(.16)	0.04(.33)	0.41(.48)	0.79(.08)
Some true	−0.02(.15)	−.015(.33)	−0.54(.41)	−0.8(.11)
Some false	0.01(.12)	0.24(.32)	0.64(.44)	0.77(.19)
Some critical (true responses)	−0.04(.17)	−0.25(.34)	−0.57(.39)	−0.8(.12)
Some critical (false responses)	−0.03(.19)	−0.18(.38)	0.21(.53)	0.78(.17)

Note: Negative values indicate movement towards the true target, and positive values indicate movement towards the false target.

Table A2.3
Experiment 3: Mean *X*-coordinates across normalized time bins.

Conditions	Bin 1 (1–25) *M*(*SD*)	Bin 2 (26–50) *M*(*SD*)	Bin 3 (51–75) *M*(*SD*)	Bin 4 (76–101) *M*(*SD*)
All false	−0.03(.2)	−0.03(.45)	0.39(.48)	0.8(.13)
All true	−0.04(.17)	−0.25(.42)	−0.6(.38)	−0.81(.11)
No critical	0.01(.2)	0.06(.42)	0.37(.45)	0.8(.13)
No false	0.0(.18)	0.11(.42)	0.47(.45)	0.8(.12)
No true	0.03(.19)	0.19(.24)	−0.13(.5)	−0.8(.12)
Some true	0.0(.18)	−0.07(.41)	−0.5(.39)	−.81(.11)
Some critical	−0.03(.22)	−0.2(.36)	0.23(.52)	0.79(.15)
Some false	−0.01(.17)	0.16(.35)	0.61(.33)	0.83(.11)

Note: Negative values indicate movement towards the true target, and positive values indicate movement towards the false target.

speaker or the calculation of alternatives to what was said. However, such a model would not necessarily require the computation of a literal meaning prior to the pragmatic meaning, as in the classical Gricean account. We look forward to the development of testable constraint-based models that implement such delays.

Conclusion

We view our research as making two major contributions to the understanding of scalar implicatures and language processing in general. First, from the point of previous research using underinformative sentences (e.g., Bott & Noveck, 2004; Bott et al., 2012; Feeney, Scrafton, Duckworth, & Handley, 2004; Noveck, 2001; Noveck & Posada, 2003; Pijnacker, Hagoort, Buitelaar, Teunisse, & Geurts, 2009), our studies reveal why upper-bound responses are delayed: participants judge underinformative sentences to be true prior to judging that they are false. The delay does not simply represent an extended period of indecision (e.g. because upper-bound responses require extra time to formulate more complicated sentence representations). Rather, participants initially formed an accessible but incorrect judgment about the truth of the sentence. Second, our results place constraints on models of how people derive implicatures: any model must predict

that people have lower-bound interpretations prior to the upper-bound. Models that compute and implement scalar implicatures in a single step are incompatible with this finding.

Acknowledgment

This work was funded by ESRC Award RES-062-23-2410 to L. Bott, T.M. Bailey and D. Grodner.

Appendix A

Categories and exemplars used in Experiments 1–3.

Exemplar	Category	Non-category	Sub-category
pigeons	birds	stones	woodpigeons
ferries	boats	mammals	P&O
novels	books	buildings	paperbacks
skyscrapers	buildings	rodents	rectangular
Ferraris	cars	clothes	convertibles
tigers	cats	fish	Indian
shirts	clothes	stones	cotton
labradors	dogs	drinks	black
poodles	dogs	vegetables	shaved
beers	drinks	snakes	lagers
wines	drinks	buildings	reds
sharks	fish	drinks	fierce
cod	fish	vehicles	smoked
roses	flowers	books	red
strawberries	fruit	flowers	wild
apples	fruit	vehicles	Coxes
beds	furniture	dogs	wooden
cockroaches	insects	flowers	German
elephants	mammals	cars	Indian
snakes	reptiles	birds	pythons
lizards	reptiles	shellfish	chameleons
rats	rodents	pens	black
lobsters	shellfish	cars	farmed
vipers	snakes	pens	poisonous
diamonds	stones	rodents	artificial
screwdrivers	tools	books	electric
elms	trees	snakes	Dutch
broccolis	vegetables	stones	purple
airplanes	vehicles	boats	Boeings
guns	weapons	cars	automatic

Note: The stimuli were constructed as subject predicate sentences according to the sentence types specified in the Methods sections.

Appendix B

See Tables A2.1–A2.3.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Bates, E., & MacWhinney, M. (1989). Functionalism and the competitive model. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of sentence processing* (pp. 3–73). New York: Cambridge University Press.

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

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

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

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

Carston, R. (1998). Informativeness, relevance and scalar implicature. In R. Carston & S. Uchida (Eds.), *Relevance theory: Applications and implications* (pp. 179–236). Amsterdam: John Benjamins.

Chierchia, G. (2006). Broaden your views: Implicatures of domain widening and the logicality of language. *Linguistic Inquiry, 37*, 535–590.

Chierchia, G. (2004). Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In A. Belletti (Ed.), *Structures and beyond* (pp. 39–103). Oxford: Oxford University Press.

Chierchia, G., Fox, D., & Spector, B. (2008). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In *Handbook of semantics*. New York: Mouton de Gruyter.

Clark, H. H., & Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology, 3*, 472–517.

Dale, R., & Duran, N. D. (2011). The cognitive dynamics of negated sentence verification. *Cognitive Science, 35*, 983–996.

Degen, J., & Tanenhaus, M. K. (2011). Making inferences: The case of scalar implicature processing. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (pp. 3299–3304). Austin, Texas: Cognitive Science Society.

Duran, N. D., Dale, R., & McNamara, D. (2010). The action dynamics of overcoming the truth. *Psychonomic Bulletin & Review, 17*, 486–491.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*, 195–225.

Elman, J. L., Hare, M., & McRae, K. (2004). Cues, constraints, and competition in sentence processing. In M. Tomasello & D. Slobin (Eds.), *Beyond nature–nature: Essays in honor of Elizabeth bates* (pp. 111–138). Mahwah, NJ: Lawrence Erlbaum Associates.

Farmer, T., Anderson, S., & Spivey, M. J. (2007). Gradiency and visual context in syntactic garden paths. *Journal of Memory and Language, 57*, 570–595.

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

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods, 42*, 226–241.

Freeman, J. B., Ambady, N., Rule, N. O., & Johnson, K. L. (2008). Will a category cue attract you? Motor output reveals dynamic competition across person construal. *Journal of Experimental Psychology: General, 137*, 673–690.

Freeman, J. B., Dale, R., & Farmer, T. A. (2011). Hand in motion reveals mind in motion. *Frontiers in Psychology, 2*, 59.

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

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

Grice, H. P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

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

Grodner, D., Gibson, E., & Watson, D. (2005). The influence of contextual contrast on syntactic processing: Evidence for strong interaction in sentence comprehension. *Cognition, 95*, 276–296.

Grodner, D., 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.

Horn, L. R. (1972). *On the semantic properties of logical operators in English*. Ph.D. thesis, University of California, Los Angeles.

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

Horn, L. R. (2006). The border wars. In Klaus von Heusinger & Ken P. Turner (Eds.), *Where semantics meets pragmatics* (pp. 21–48). Oxford: Elsevier.

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

Just, M. A., & Carpenter, P. A. (1980). A theory of reading – From eye fixations to comprehension. *Psychological Review, 87*, 329–354.

Levinson, S. C. (2000). *Presumptive meanings*. Cambridge, Mass.: MIT Press.

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review, 101*, 676–703.

McElree, B. (1993). The locus of lexical preference effects in sentence comprehension: A time-course analysis. *Journal of Memory and Language, 32*, 536–571.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling thematic fit (and other constraints) within an integration competition framework. *Journal of Memory and Language, 38*, 283–312.

Noveck, I. (2001). When children are more logical than adults. Experimental investigations of scalar processing costs in implicature production. *Cognition, 78*, 165–188.

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

Noveck, I., & Sperber, D. (2007). The why and how of experimental pragmatics: The case of scalar inferences. In N. Roberts (Ed.), *Advances in pragmatics*. Basingstoke: Palgrave.

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

Reed, A. V. (1973). Speed–accuracy trade-off in recognition memory. *Science, 181*, 574–576.

Rips, L. (1975). Quantification and semantic memory. *Cognitive Psychology, 7*, 307–340.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In D. E. Rumelhart & J. L. McClelland (Eds.). *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1). Cambridge, MA: MIT Press.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (2nd ed.). Oxford: Blackwell.

Spivey, M. J. (2007). *The continuity of mind*. New York: Oxford University Press.

Spivey, M., Grosjean, M., & Knoblich, G. (2005). Continuous attraction toward phonological competitors: Thinking with your hands. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 10393–10398.

Tabor, W., & Tanenhaus, M. K. (1999). Dynamical models of sentence processing. *Cognitive Science, 23*, 491–515.

Traxler, M. J., Pickering, M. J., & Clifton, C. Jr., (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language, 39*, 558–592.

Urbach, T. P., & Kutas, M. (2010). Quantifiers more or less quantify on-line: ERP evidence for partial incremental interpretation. *Journal of Memory and Language, 63*, 158–179.

van Gompel, R. P. G., Pickering, M. J., Pearson, J., & Liversedge, S. P. (2005). Evidence against competition during syntactic ambiguity resolution. *Journal of Memory and Language, 52*, 284–307.