Scalar implicatures in a Gricean cognitive system

Richard Breheny, University College London

{Draft - To appear in, N. Katsos & C. Cummins, (eds) *Handbook of Experimental Pragmatics*. Oxford University Press. Please do not quote or copy}

1. Introduction

1.1. Overview

This paper reviews recent experimental research into questions about how language and other functions of the mind are integrated when humans communicate. I posit a Gricean system that serves this purpose and discuss how recent developmental and ethological research provides evidence for such a system's existence. Subsequently we focus on the much-studied phenomenon of scalar implicature. We first consider the phenomenon of scalar implicature in the broader context of pragmatic effects. A short review of theoretical debates as to the status of various sub-types of scalar phenomenon is followed by sections that discuss experimental research relevant to different interfaces in the Gricean system when it comes to scalars.

1.2. Pragmatics, Communication and a Gricean System

Although it hardly began this way, the field of Pragmatics is now increasingly seen as a discipline that is the concern of the psychological sciences. From a psychological perspective, one can consider the systems or functions that the mind realises in terms of domains such as vision, language, memory, reasoning and so forth. Pragmatics deals with communication as a domain that is somewhat separate from language. Pragmatics typically views language as a function in the generative sense - as something which computes phonological and syntactic representations, as well as providing sets of rules for composing meanings of constituents. From this perspective one role of pragmatics is to characterise the systems for computing a mapping between the thoughts we express in situated linguistic utterances and the linguistic structures used in those utterances. A more general role is to explain human communication, be it linguistic or non-verbal.

There is a growing consensus that human communication is very distinctive from that of other primates or other species in its use of social-cognitive abilities (Tomasello et al., 2005; Herman et al., 2007). Research on infants and primates reveals that humans are distinct in assuming that communicative signals create shared information (Tomasello et al., 2005; Moll & Meltzoff, 2011; Csibra, 2010), and that the signaller's aim is to provide a kind of *good* (Tomasello 2008 Csibra &

Gergely, 2011). Moreover, infants have been shown to display sensitivity to unseen mental states of communicators - their intentions, beliefs and desires - in the earliest stages of communicative development (Southgate et al., 2010; Liebal et al., 2010). Infants also demonstrate expectations that other agents are rational in achieving their aims (Gergely & Csibra, 2003).

It comes as no surprise then to find that the ideas of Grice and his contemporaries (Grice, 1957; 1975, Lewis, 1969) are still current in the field of pragmatics. A 'Gricean' account of some apparently linguistic phenomenon typically attempts to explain attested meanings, in part, by appeal to a system for deriving meaning that a speaker intends their communicative gesture, or utterance, to have. In a typical Gricean explanation, the speaker's meaning goes beyond, or departs from, the meaning which can be recovered from any information encoded in the stimulus via a lexicon and grammatical rules. Gricean explanations take for granted a commonly shared level of rationality plus some common expectations about what good the communicator aims to provide. Gricean accounts explain our intuitions about utterance meaning in terms of inferences about which thoughts a speaker intends to share. From a cognitive perspective, if we follow Grice's proposals about how to account for certain aspects of what a speaker means, we can imagine an inferential system that interfaces a range of cognitive functions to make decisions about speaker's intended meaning. These functions include language, Theory of Mind and long-term memory/conceptual structure.

Given the developmental research, then, human infants appear to possess abilities that would constitute a Gricean system for inferring speaker's meaning. That is, they have a cognitive system that integrates information about the linguistic stimulus and the utterance situation, including the expectations and goals of the interlocutors, to compute the thoughts that the speaker intends to share. Moreover, the above research suggests that infants possess these abilities at a point where they barely manifest linguistic abilities beyond the one-word stage. Thus we should be encouraged that a Gricean system realises a naturally occurring function that distinguishes human cognition from that of other species.

Scalar Implicature is one domain where linguists and philosophers have proposed Gricean accounts of apparently linguistic phenomena. In this chapter, 'Scalar Implicature' refers to a broad class that includes the sub-categories, 'Straight Scalars' (SS), 'Ignorance Inferences' (II) and 'Embedded (Scalar) Enrichments' (EE). These will be illustrated in Section 2.1 below, and contrasted with other pragmatic effects. A diverse range of broadly Gricean accounts of scalar phenomena have been offered over the years (Grice, 1975; Horn, 1972; Gazdar, 1979; Soames, 1982; Levinson, 1983; 2000; Sperber & Wilson, 1995; Benz & van Rooij, 2006; Schulz & van Rooij, 2006; Franke, 2011; Russell, 2012; Frank & Goodman 2012; among others). These Gricean accounts have not gone unchallenged (Chierchia, 2004; Fox, 2007; Chierchia et al., 2012; a.o.) due to some well-known

problems that arise. This debate will briefly be rehearsed in section 2.2. The issue here has come to be about how language and Gricean systems interface.

Experimental research in pragmatics has addressed, or is relevant to, this language/pragmatics interface question when it comes to Scalar Implicatures. Are they a purely extra-grammatical phenomenon? Are scales stored in a mental lexicon? Are there default scalar implicatures? Are some scalar phenomena better described solely within a linguistic framework? Note that the status question applies independently to SSs, IIs and EEs. For each sub-phenomenon, there are many nuanced shades in the theoretical positions adopted. In Section 3, we will review experimental research which can contribute to these debates.

To the extent that SI phenomena are not purely linguistic, we can ask how are the results of linguistic and extra-linguistic computations integrated in a Gricean inferential system? Experimental research has also produced results that bear on this question. This is reviewed in Section 4.

2. Scalar Implicature

2.1. Domain of Inquiry

Scalar Implicature is a term that has been used to refer to a wide range of phenomena. As will be discussed briefly below, there is considerable debate about how best to account for these phenomena. Under most current approaches scalar phenomena do not get a unitary analysis. Moreover, different, opposing accounts have been offered for the same sub-category of Scalar Implicature. In order to not pre-judge the nature of these phenomena, it is best give an initial sketch of the domain of the research to be discussed here by prototypical example.

Let us begin some examples of what is the most commonly discussed kind of case - what I shall call the 'Straight Scalar' (SS). Consider (1-2), where what follows '~>' would be a plausible implication in easily imaginable situations:

1. Some of the students got an A on the test.
   ~> Not all of the students got an A on the test.
2. A: Jane was planning to cut the grass and wash the car. How did she get on?
   B: She cut the grass.
   ~> Jane did not wash the car.

The Straight Scalar implication has been characterised as the negation of some proposition that is related to the assertion in virtue of lexical association, as in (1), or in virtue of contextual salience or

relevance as in (2), and perhaps also (1). The negated element is almost universally referred to as an *Alternative*. For the implication under (1), the alternative proposition would be *All of the students got an A on the test*. For (2), it could be *Jane cut the grass and washed the car* or simply, *Jane washed the car* – depending on one's theory of scalars (see Katzir, 2007).

(1) is an example where there appears to be a lexical association involved in the relation between the asserted proposition and the alternative (in virtue of the fact that the alternative could be derived via a replacement of 'some' with 'all'). Here we will refer to such cases as lexical scalars. More frequently perhaps, alternatives become available in virtue of context alone, as in (2). Here, as elsewhere, these examples will be referred to as *ad hoc* scalars.

The next sub-category of scalar phenomena is the 'Ignorance Inference'. Two examples are given below, both of which are discussed in Grice's William James Lectures (see Grice, 1975; 1978).

3. A is planning with B an itinerary for a holiday in France. Both know that A wants to see his friend C, if to do so would not involve too great a prolongation of his journey.
   A: Where does C live?
   B: Somewhere in the South of France
   ~> The speaker does not know where in the South of France

4. A: Where is Max planning to visit when he goes to Scotland?
   B: Edinburgh or Glasgow.
   ~>The speaker is not certain that Max is planning to visit Edinburgh
   ~>The speaker is not certain that Max is planning to visit Glasgow

In both cases, there is an implication that the speaker does not know about some contextually related propositions. The similarity between Ignorance Inferences and Straight Scalars can be made to appear stronger if we take the Gricean perspective on straight scalar implicatures (see below). That is, we assume that a speaker who implies *Not all of the students got an A* does so in virtue of communicating their belief that not all of the students got an A. In other words, where the alternative *A* is negated in SS, it is in virtue of the speaker conveying that they believe *not A*. By contrast, the speaker conveys something weaker in II, which is that they do not have the belief that A. In both cases, arguably, an alternative (*A*) is involved.

The final sub-category to be discussed here is Embedded Enrichment. One way to characterise embedded scalar enrichments of an assertion is to imagine that a sub-constituent of the assertion has been enriched by SS and the resulting enriched sub-constituent contributes to the

overall proposition being expressed. For instance, a reading of (5a) containing an Embedded Enrichment could be glossed by imagining the constituent 'hit some of the targets' being given a reading, *hit some and not all of the targets*. This is indicated in (5b).

5. a. Exactly one player hit some of the targets.
   b. Exactly one player hit some but not all of the targets

This example is taken from Potts et al. (2015) which reports that participants in an experiment readily understood sentence (5a) according to the gloss in (5b). The significance of this sub-category of scalar phenomena will be discussed below. Here we can note that, unlike with SS or II, it is not possible in many cases of EE to describe the enrichment of the linguistically determined meaning in terms of the conjunction of the literal meaning of the sentence used and some other proposition. For instance in (5), the literal meaning of (5a) and the enriched meaning, (5b), do not stand in an entailment relation with each other.

This ends the survey of core phenomena typically discussed under the heading of Scalar Implicature. I round up this section with some comments on the location of these exemplars in the domain of pragmatics more generally.

Above I mentioned some prototypical examples of scalar phenomena. Over the years, linguists and philosophers have argued about other kinds of examples, as to whether or not they are cases of scalar phenomena. Below I list three cases which will be relevant when we come to discuss experimental work on scalars:

6. Mary has two children.
   ~> Mary has no more than two children.
7. I will give you $5 if you wash my car.
   ~> I will give you $5 only if you wash my car.
8. You can have coffee or cake.
   ~> You can have coffee and you can have cake.

A straight scalar approach to cases involving numerals (Horn, 1972; Gazdar, 1979; Levinson, 1983), proposes that 'two children', like 'some children', has an encoded meaning that could be glossed in purely lower-bounding terms – something like *two or more children*. Thus (6) would be true even if Mary has three or more children. The implication that she has no more than two is considered to be a Straight Scalar, with the alternative being *Mary has three (or more) children*. This view has been

contested, however (see Horn, 1992; Geurts, 2006; Breheny, 2008), and the issue remains somewhat open.

In contrast to the numeral case, early Gricean accounts of so-called 'conditional perfection' argued against treating the implication indicated under (7) as SS (see Atlas & Levinson 1981), but treated it as another kind of implicature. However, others disagree and would treat this, or related implications as SS (Matsumoto 1995; von Fintel, 2001; Franke, 2009; Geurts, 2010; Horn, 2000; Levinson, 2000).

(8) is an illustration of a so-called 'free choice' inference. Again, there is not universal agreement about how to explain the implication illustrated here. Some proposals favour an account based purely on the linguistic expressions used (Barker, 2010; Geurts, 2005; Zimmermann, 2000). However, many recent discussions propose that free-choice inferences can be explained by applications of the mechanism for straight scalars (Klinedinst, 2007; Fox, 2007; Franke, 2011; Geurts, 2010).

Beyond the domain of scalars, a great number of phenomena have been classed as 'implicature' or have been analysed in broadly Gricean terms. These include Relevance implicatures (Grice, 1969a), politeness implicatures (Brown & Levinson, 1987), figurative language such as metaphor, hyperbole, metonymy and irony (Grice, 1969a; Sperber & Wilson, 1986), loose use or approximation (Sperber & Wilson, 1986), M-(anner) implicatures (Levinson, 2000; Sperber & Wilson, 1986) and other enrichments variously described as I-implicatures or R-implicatures (Levinson, 1984, 2000; Horn, 1984; Carston, 2002). Also, diagonalisation accounts of existential closure (Stalnaker, 1979), and certain accounts of presupposition (Stalnaker 1974; Simons, 2001) and anti-presupposition (Sauerland, 2008) have been cast in broadly Gricean terms. While space considerations preclude us looking at these other phenomena in detail, it is relevant to the discussion of experimental work on scalars to consider a broad class of pragmatic effects that contrast in interesting ways with scalars. This could be called the class of R-/I-/Relevance implicatures. Two examples are given below:

9. A comes upon B, who is looking under the bonnet of an obviously immobilised car.
   A: There's a gas station around the corner
   ~> The speaker believes it likely that the gas station is open

10. Mary tripped on a Persian carpet and she broke her wrist.
    ~> the wrist break resulted from the tripping over

These implicatures contrast with scalars insofar as they result in the speaker communicating a stronger salient proposition (*there's an open garage around the corner*; *Mary tripped on a Persian rug and she broke her wrist as a result*), rather than the negation of a stronger salient proposition. Also, arguably, they do not involve alternatives (Carston, 2002, but see Poppels & Levy, 2015).

In addition to implicature and related phenomena where literal or linguistic meaning is either added to or changed, many have argued that a Gricean system plays a key role in decisions about what expressions used in an utterance literally mean (Wilson & Sperber, 1981; Neale 1992, Kehler & Rohde, 2016). A typical illustration of this is disambiguation. Words can be ambiguous due to having multiple meanings and strings of words can be ambiguous due to the possibility of multiple underlying syntactic structures. When utterances are produced in context, they typically pose a decision problem as to the intended parse. To the extent that a Gricean system is employed in this decision process, then this also impacts on experimental research on scalars that seeks to detect the presence of Gricean vs. Linguistic computations, as we will see below.

Although a formal corpus survey of the whole domain of linguistic pragmatics has yet to be undertaken, informally it seems that instances of scalar phenomena are very much in the minority when it comes to the domain of pragmatics. What has been more rigorously established (Degen, 2015), is that even for the prototypical lexical-scalar trigger, 'some', actual tokens found in the corpus were judged to give rise to a Straight Scalar implicature (as illustrated in (1)) less than half of the time. In addition, van Tiel et al. (2016) show that, if anything, other scalar terms would give rise to scalar implicature at lower rates than 'some'. And yet, although it is perhaps a minority phenomenon, and not as prevalent as one might think even for lexical triggers, it has attracted more attention than any other in recent pragmatics research, particularly in experimental pragmatics.

2.2. **Scalars as Gricean phenomena?**

Grice's William James Lectures (Grice, 1975; 1978) contain accounts of two examples of Ignorance Inferences. These are given in (3-4) above. In his programmatic 'Logic and Conversation' Grice reserves a special category of conversational implicature for this phenomenon. This is Group B, where a speaker violates the maxim of informativeness on account of her desire not to violate the maxim of truthfulness. Group B implicatures result in an ignorance inference – that the speaker does not provide the required information because to do so would be to assert something for which they lack adequate evidence. These texts by Grice contain no clear discussion of Straight Scalars. However, subsequent work in the Gricean vein locate Straight Scalars within Grice's Group A of pragmatic phenomena (Horn, 1972; Gazdar, 1979; Soames, 1982; Sauerland, 2004; Geurts, 2010). Here the speaker merely exploits the maxims of informativeness and truthfulness in order to get

across the required information implicitly: The speaker is judged to have not explicitly expressed a more informative alternative to their assertion on the grounds not of ignorance but of believing that alternative to be false.

Before we turn to Gricean approaches to EE, we can consider other, broadly Gricean approaches to scalars. Recent Rationalist, Bayesian models of Scalars in the Game Theory tradition (Benz & van Rooij, 2006; Franke, 2012; Frank & Goodman, 2012) focus on the interest that speaker and hearer have in co-ordinating beliefs about the world. In Rational Speech Act approaches to scalars, as articulated in (Frank & Goodman, 2012) and (Bergen et al., 2016), informativeness comes into the derivation through a notion of specificity – an optimal speaker uses the most specific utterance they can, relative to what they believe about the world. Thus, if one makes assumptions about the alternatives we reason over (see Frank & Goodman, 2012), a speaker who utters (1) when they could have used 'all' is judged to be less likely to believe *all* than *some and not all*. Alternatively, *Relevance-based* approaches to scalars, some of which operate within the probabilistic/Bayesian tradition (Sperber & Wilson, 1995; Russell, 2012; Merin, 1999) work with a comparative notion of relevance and propose that speaker and hearer reason about each other on the basis that the utterance is relevant. While the asserted utterance must be presented as at least relevant to some degree, alternatives may be judged as potentially more relevant, leading to either SS or II, depending on what can be assumed about the beliefs of the speaker. One observation made within relevance-based frameworks (see in particular, Russell, 2012; Sperber & Wilson, 1995; but also Magri, 2009) is that while a straight scalar implicature may make the utterance more relevant, the explicitly asserted proposition has to be at least adequately relevant in the context. Russell (2012) illustrates this point with the following kinds of example:

11. a. #Oh crap! Some of the students passed
    b. Oh crap! Only some of the students passed
    c. Oh crap! Not all of the students passed

In the context of an interjection like, 'Oh crap!', we can assume that (11b,c) are felicitous due to the fact that *not all* is part of the explicit assertion, as it provides the relevant bad news. The infelicity of (11a) suggests that it is not sufficient for the speaker to merely scalar-implicate *not all* to achieve a basic level of relevance in context. Evidence such as this suggests a difference in status of the *some* and *not all* implications between (11a) and (11b). This is a point that bears on experimental research, as discussed below.

Turning to EE, as with SS, Grice makes no proposals in Logic & Conversation for these. In fact, it has been a long-standing criticism of 'Logic and Conversation' and many subsequent Gricean proposals that they cannot account for many Embedded Enrichment effects (Cohen, 1971; Wilson, 1975, Carston, 1988). There is however scope within the general Gricean program to deal with local adjustments to meaning (Grice, 1957; Sperber & Wilson, 1986; Carston, 1988; Neale, 1993). In the Relevance Theory tradition, it is assumed that there is no presumption of literal truthfulness among language users, only of Relevance (in the technical sense of RT). Translated into psychological terms, the proposal is that the lexically encoded association between expressions and meanings serves only as a starting point when it comes to inferring the proposition explicitly expressed by an utterance. The theoretical question of some interest then concerns the mechanisms that are responsible for modulating lexically encoded meaning. Recent proposals in the RSA framework (Bergen et al., 2016; Potts et al., 2015) also implement the idea that the computation of speaker meaning does not presume literal truthfulness. Rather, speakers and hearers take for granted that linguistic expressions may be mapped to novel meanings in context. Uncertainty about which mapping the speaker intends then becomes part of the bigger decision problem, mentioned above, concerning under-specification of explicit meaning, given the linguistic form used in the utterance.

Notwithstanding the generality of Embedded Enrichment phenomena and the fact that a formal description of the phenomenon is possible within a broadly Generative framework, recent *non-Gricean*, 'Grammatical' approaches to scalars have cited EE as grounds for bringing both SS and EE phenomena back within a more squarely linguistic description. Grammatical Theory proposals are outlined in Chierchia et al., 2011; Fox, 2007 and elsewhere. The basic idea is that SS can be represented in the syntactic structure for an utterance via a covert exhaustification operator, *exh*, that functions much like 'only'. Thus (1), repeated in (12a) below, could have a syntactic representation (LF) as in (12b), with *exh[S]*, defined relative to a set, *Alt*, of alternative sentences in (12c)

12. a. Some of the students got an A on the test.
    b. [exh [Some of the students got an A on the test.]]
    c. $||\text{exh}_{\text{Alt}}[S]|| = 1$ iff $||S|| = 1$ and for all $A \in \text{Alt}$, unless $||A|| \subseteq ||S||$, then $||A|| = 0$

Embedded scalars then become straightforward to accommodate, given that *exh* can appear at scope sites other than the root clause. For example, an LF for (5a) could be as in (13):

13. [Exactly one player$_i$ [exh$_{ALT}$ [t_i hit some of the shots]]]

Aside from EE phenomena, an important motivation for the GT comes from a problem for the general Gricean programme that has yet to be mentioned – the Symmetry Problem. This is a problem about the choice of alternatives for SS (and perhaps for EE). It is widely recognised that accounts based on Grice's proposals cannot explain why speaker and hearer would reason about an alternative *A*, when its 'symmetric' alternative *not A* should be an equally viable alternative (see Fox, 2007; Katzir, 2007 for details). It seems something other than informativeness and relevance to a purpose impact on the derivation. Grammatical theorists would argue that purely linguistic, or structural, factors are involved in the computation of alternatives (Fox & Katzir, 2011; Katzir, 2015 – but see Romoli, 2013; Trinh & Haida, 2015; Breheny, Klinedinst, Romoli & Sudo, *forthcoming* for a discussion of well-known problems). If this conjecture is correct, it motivates a view of SS and EE as more squarely linguistic phenomena. It should be noted, however, that even the structural approach to alternatives presumes that Gricean systems need to interface with linguistic systems when it comes to selecting the contextually relevant alternatives from among the formally defined alternatives.

3. Experimental research at the language-Gricean interface

3.1. Default implicatures: A testable proposal about the language-pragmatics interface

Levinson (2000) developed a hybrid Gricean account of scalar phenomena which includes the idea that at least lexical scalar implicatures have a special default status, lying between purely linguistic phenomena and purely Gricean pragmatic phenomena. Levinson conjectures that being default in this way has processing implications in that lexical scalar items trigger the activation of the scalar implication and only a strong contextual bias against this implication would result in its suppression. To give an example, the implication indicated under (1) would be a default, and so we would expect comprehenders to activate that implication irrespective of context. Note that, according to all approaches, including Levinson's, a sentence containing a lexical trigger like 'some' may or may not be understood to carry the straight scalar implicature and, as mentioned, a corpus survey shows that it does so slightly less than half the time (Degen, 2015). Thus, where an utterance is understood to not carry the implication, the default account must assume some process of de-activation or suppression.

Experimental research tested Levinson's prediction that participants who include a scalar implicature in their response should do so either faster or at least not slower than participants who

do not. This was tested in Bott & Noveck (2004). Participants were given a verification task based on world knowledge. Critical items asked participants to verify items as in (14a,b):

14. a. Some anchovies are fish
    b. Some fish are anchovies

A participant would judge (14a) true when the straight scalar implicature is not part of their interpretation of the item, but false when the scalar implicature, *Not all anchovies are fish* is part of the understanding. By contrast, (14b) is true with or without the implicature. Bott & Noveck report that with-implicature responders had more errors and took longer than no-implicature responders on critical items like (14a). The results run contrary to Levinson's predictions. They also showed that a group of participants given less time to respond were more likely to give a no-implicature response than a group given more time. Noveck & Posada (2003) report similar results.

Moving from sentence verification to reading-time studies, Breheny et al. (2006; see also Katsos et al. 2006) demonstrate that participants take longer to read segments carrying implicature triggers ('some of the consultants') when reading texts where context biases the straight scalar implicature compared to texts where they do not. This result was replicated in Bergen & Grodner (2013) with items that controlled for a repeated-name penalty that might otherwise explain the previous reading-time results (but see Politzer-Ahles & Fiorentino, 2013).

Further independent disconfirmation of the default position comes from Speed-Accuracy trade off (SAT) studies reported in Bott et al. (2012). Using verification tasks similar to Bott & Noveck (2004), Bott & colleagues report that responses emerged faster for the no-implicature group than the with-implicature group. Finally, using a mouse-tracking paradigm, Tomlinson et al. (2013) report that participants who respond 'False' to an item like (14a) first move the mouse cursor in the direction of a 'True' response before correcting. By contrast, 'True' responders move the mouse directly to the 'True' side of the screen. Levinson's default approach predicts that, if anything, the opposite pattern would be observed.

Taken together, these results undermine the default position, as that was developed by Levinson. This being the case, do these results also support a broadly Gricean position, as against, say, the Grammatical Theory? We turn to this question in the next section.

3.2. **Counting the cost of scalar implicatures**

One recurrent theme emerging from research on scalars is that a with-scalar response takes longer than a without-scalar response, although this is not a universally reported finding (see for instance

Feeney et al., 2004; Politzer-Ahles & Fiorentino, 2013; and references cited in section 4.1 below). Evidence for cost or delay for scalars has been reported in verification tasks (Bott & Noveck, 2004; Bott et al., 2012), Reaction time reading (Breheny et al., 2006; Bergen & Grodner, 2012) and visual-world eye-tracking (Huang & Snedeker, 2009). These results seem supportive of the contextualist idea that computing a Straight Scalar implication involves enriching the encoded non-SS meaning in context. According to Gricean approaches, Straight Scalars and Ignorance Inferences involve reasoning about something that a speaker could have said but didn't, and further coming up with a reason for the speaker not giving more information, in terms of what the speaker was in a position to know. In the literature, the first part of this process is discussed in terms of finding contextually relevant alternatives. The second part is referred to as an 'epistemic step' in the case of SS, although a similar process must take place in the case of Ignorance Inferences. It may seem that, by contrast, the Grammatical account of SS short-circuits these stages by allowing that an operator be inserted into the syntactic structure for the sentence. However, matters may not be so straightforward. Recall that it is widely argued that decisions relating to facets of the linguistically determined proposition in context ultimately fall within the purview of Gricean pragmatics. In particular, a process of contextually determined domain restriction (or reference assignment) is involved in virtually all noun phrases uttered. The process of deciding on the set of alternatives for the exhaustivity operator is virtually identical to the process of choosing a domain for the operator, 'only'. That is, both *exh* in (12b) above and its explicit cousin, 'only' quantify negatively over alternative propositions. Thus, the presence of an exhaustification operator in a syntactic structure entails a stage of selecting alternatives, more or less equivalent to the alternative-selection stage in the Gricean derivation. Moreover, even the decision to insert an operator or not to do so would depend on similar factors. Thus, a Gricean process of alternative selection may be required both for the derivation of a SS as a Gricean inference, and its derivation via a covert linguistic operator.

This still leaves a potential difference between Gricean and Grammatical derivation of Straight Scalars since the Gricean derivation mandatorily includes a step which considers the speaker's reasons for not asserting alternative propositions. Experimental research that manipulates whether the speaker would know that the alternative were true or not has shown that participants are sensitive to this step. In both Bergen & Grodner (2012) and Breheny et al. (2013) effects of incorporation of straight scalar implicatures disappear in conditions where the speaker is ignorant of the facts. This sensitivity to an 'epistemic step' is not strongly predicted by Grammatical Theory. However, these studies are not completely decisive since the Grammatical Theory does endorse the Gricean account of ignorance inferences (see Fox, 2007 in particular). Thus, the results in both Bergen & Grodner (2012) and Breheny et al. (2013b) could be accounted for in terms of the salient

ignorance of the speaker in the experimental contexts. A Gricean system that decides on the inclusion or not of a covert operator would withhold inclusion in ignorance contexts, favouring instead an ignorance inference.

One straightforward means to test whether the costs of SS simply boil down to the costs associated with the insertion of a covert, *exh* operator would be to compare cases when participants incorporate SS with cases that involve sentences with an overt exhaustification operator, such as 'only'. This comparison has been made several times in the experimental literature and the results are informative. In the SAT research reported in Bott et al. (2012), one experiment has participants responding pragmatically to items like in (14a) above and includes items that are just like (14a) but for the inclusion of 'only': 'Only some anchovies are fish'. Bott and colleagues report that participants depart from random response sooner in the 'only some' condition than the pragmatic 'some' condition. Given that the function denoted by *[only [some anchovies]* and *[exh [some anchovies]]* would involve the same levels of complexity to compute and verify, it seems that there is some extra cost to processing the SS items compared to their explicit counterparts. This is not predicted by the Grammatical Theory. By contrast, Gricean approaches could argue that a richer representation of context is entailed by SS than by 'only some'. That is to say, a Gricean scalar implicature involves representing a speaker's reasons for not using the alternative and this is somehow the cause of the delay.

A comparison between straight scalar items like (14a) and 'only some' items has also been made in a paradigm that manipulates working memory load. The paradigm was introduced to experimental pragmatics in de Neys & Schaeken (2007), who demonstrated that participants gave fewer SS-based responses to items like (14a) when a parallel task made greater demands on working memory than in a low memory-load condition. Marty & Chemla (2013a) follow up on this work by comparing the effects of memory load in a 'some' condition with an 'only some' condition. Their results in the 'some' condition replicate de Neys & Schaeken (2007). They also report that while rates of with-SS responses drop in the high load case, this does not occur in the explicit 'only some' case.

It is possible to interpret this result as favouring a Gricean account of SS over the Grammatical Theory. The idea would be that the Grammatical Theory implies that the process of representing and verifying the items in both the implicit and explicit conditions should come with comparable costs. Thus no difference would be predicted. By contrast, as mentioned, Gricean accounts tend to entail a richer representation of the utterance situation for SS items than 'only some' items. This would explain the greater effect of memory load in the with-SS condition.

However, Marty & Chemla caution against drawing this conclusion since the Grammatical view entails an extra layer of complexity for the with-SS condition, compared to the 'only some' case. This is due to the fact that, according to the GT, the sentence in (14a) is inherently ambiguous between a parse that includes the exhaustivity operator, *exh*, and one which does not. Marty & Chemla argue that it is possible that the extra layer of decision-making involved in the disambiguation makes access to the implicature susceptible to extra memory load.

This alternative account awaits confirmation but it is possible to make some observations based on results that we have already. First, ambiguity does not always imply extra cost. For example, Duffy et al. (1988) show that reading times for equi-biased homonyms (i.e. ambiguous words whose senses are equally frequently employed) do not differ from unambiguous controls when the target stimulus is supported by context. By contrast, out of context, there is a cost. Returning to the corpus survey in Degen (2015), we can observe that around 50% of samples containing 'some' were judged to give rise to an implicature. If we assume scalar implicature is represented in linguistic structure, we seem to have a case of equi-biased ambiguity. The question then comes down to whether the items are supported by context or not. In lexical disambiguation research, an ambiguous word appears only once in an experimental session. So context for that word would be associated closely with the trial in which it appears. In the memory load studies, the same target expression appears numerous times in virtually the same context. Thus, if anything, we should consider these items to have some supporting context, especially for participants who tend to respond one way or the other across the bulk of trials (see Bott & Noveck, 2004; Feeney et al., 2004; Antoniou et al., 2016 among others). Thus, if we assume that the SS ambiguity is equi-biased but the experimental setting provides supporting context for the parse that a participant gives, we should not expect these 'some' trials to engender extra cost in virtue of any ambiguity inherent in the stimulus.

We could, on the other hand, assume an equi-biased ambiguity but a neutral context in these working memory studies. If that were the case, it would be instructive to compare the effects of memory load on latencies of responses that include the implicature to those that do not. This comparison is made in de Neys & Schaeken (2007; see also Dieussaert et al., 2011). They report that 'False' responses (i.e. with-SS) are delayed under high load compared to low load. By contrast, 'True' responses (i.e. with no SS) are not delayed in the higher load condition. If it is an equi-biased ambiguity of 'Some fish are anchovies' that gives rise to a different result under load, then we should expect a similar difference regardless of whether participants choose the with-SS response or the without-SS response. Thus, the GT account is not confirmed by de Neys & Schaeken's results. Nor is it confirmed by a comparison of reaction times for filler items, including those like (14b) above.

Again, we should expect to see a difference under load for a set of items that contain ambiguous 'some', but no difference is found.

An alternative explanation for the memory-load comparison could appeal to the fact mentioned in Section 1.2, that even where a Straight Scalar is derived, the asserted, lower-bounded meaning of the utterance should be relevant in its own right. In cognitive terms, we can assume that this means that where items like (14a) give rise to an implicature, the information that some or all anchovies are fish should attract attention in its own right. The negative proposition that not all anchovies are fish, being implicated, should not be more salient than the assertion. This idea is supported not only by examples like those in (11), taken from Russell (2012), but also by the mouse-tracking studies reported in Tomlinson et al (2013), which show that participants who respond 'False' to (14a) are drawn first to a 'True' response (consistent with the lower-bound meaning of the asserted sentence), then shift mouse direction.

By contrast, the corresponding sentence with 'only' ('Only some anchovies are fish') does not prioritise the lower-bounded component in this way. In fact, there is a broad consensus that the semantic, asserted content is *not all anchovies are fish*, while there are arguments that the positive, lower bounded part of the sentence is either a scalar implicature or a presupposition (Horn, 1992; van Rooij & Schulz, 2007; Ippolito, 2007). If that is right, then the positive part of the 'only' item could not be more salient than the negative.

How would an 'explicit relevance' account of these memory load data go? Let us assume that what is explicit and (assumed to be) relevant must receive attention, in virtue of being activated by the linguistic stimulus. By contrast, potential scalar implicatures need not receive attention in virtue of the linguistic processing of the stimulus itself. Then it stands to reason that scalar implicatures are more likely to not be activated under increased memory load than asserted parts. For the 'some' items, this means that the negative, 'not all' component is liable to be not activated in high-load conditions, while for the 'only some' items, if anything the lower-bounding 'some or all' component is susceptible. In the latter case, not accessing the positive component would not affect rates of 'False' judgements on critical items, since that judgment is based on the negative, *not all* component being activated (see also Tomlinson et al., 2013; Antoniou et al., 2016). By making these assumptions about the effect of a requirement that the explicit content is relevant, we can explain the pattern of results in this kind of memory-load study.

This explanation of the working-memory effect and other effects (such as the mouse-tracking effect) is consistent with constraint-based accounts of processing scalar implicatures (Grodner et al., 2010; Degen & Tanenhaus, 2015), which argue that even comprehension processes

that require rich contexts need not be delayed or costly, if the relevant information is sufficiently activated or available. We return to a more detailed discussion of the processing question below.

3.3. Experimental research on potential scalars

Until now, we have considered what experimental research tells us about the status of Straight Scalar implicatures, whether they should be viewed as 'Gricean' or otherwise. As mentioned in section 2.1, there are several phenomena that attract debate as to their status – whether they should be classified as scalar implicature in the first place or explained in some other way. In this section, I will discuss recent experimental research that bears on these debates.

Let us begin with the case of numerically quantified noun phrases. The example in (6), repeated below, indicates that sentences containing noun phrases like, 'two children', can (and often do) carry an upper-bounding, 'no more than two' implication:

(6) Mary has two children.
~> Mary has no more than two children.

This could be explained as a straight scalar implicature if the sentence containing, 'two children' encodes a lower-bounding meaning, *two or more children* and where the alternative for SS is formed by replacing the numeral, '2', with '3'. However, a number of differences between straight scalar items like, 'some' and numerals have been observed (Horn, 1992; Breheny, 2008), leading to scepticism that the implication under (6) is in fact a scalar implicature. Many have argued that the implication under (6) is an entailment of an encoded meaning for the sentence (Geurts, 2006; Breheny, 2008; Kennedy, 2015, Buccola & Spector, 2016). This would be the case if the noun phrase is understood literally to mean something like *exactly two children* – although there are other analytical tools to account for this entailment (see Kennedy, 2015; Buccola & Spector, 2016). There are, however, many occasions when numerically quantified noun phrases have a lower-bounded-only reading. For example, in (15) below, the sense of 'Mary must have two children' is better glossed as 'Mary must have two or more children' rather than, 'Mary must have exactly two children':

15. To receive this benefit, Mary must have two children.

Different theories have different approaches to these data. Some posit an ambiguity in the sentence between an *exactly two* and a *two or more* reading (see Geurts, 2006, Kennedy, 2015; Buccola &

Spector, 2016). Breheny (2008) takes a different route, arguing that a separate pragmatic mechanism operates to make available the *two or more* reading in this case. Both kinds of approach get some support from a recent study that looks at modified numerical expressions, as in (16):

16. a. Less than three dots are red.
    b. Between two and five dots are red.

It happens that the mechanism for deriving the lower-bounded reading in (15) also derives an existential reading of (16a,b) meaning that either of these sentences could be truthfully uttered of a situation where seven dots are red. Marty et al. (2015) report that participants asked to verify (16b) for displays involving seven red dots, respond 'True' in such a situation at a higher rate than their 'False' response to situations that make (16b) false according to either lower-bound-only or doubly bound readings (i.e. where only one dot or no dot is red). This data points to participants seeing 'between m and n' sentences as having a lower bounding reading, as predicted by these non-SI accounts of numerical expressions.

More relevant data that numerically quantified noun phrases are not like other scalars comes from Marty et al. (2013) which employs the memory load paradigm to compare the status of the readings of 'two dots' versus 'some dots'. Recall that, under higher memory load, participants less frequently include the upper bounding *not all* inference in verifying sentences with 'some'. This means that more responses based on the lower-bound reading *some or all* are elicited under high memory load (de Neys & Schaeken, 2007; Marty & Chemla, 2013). Marty et al. (2013) replicate this effect for target items, 'Some dots are red' when the visual display shows only red dots. However, for target items, 'Four dots are red' with a screen displaying six red dots, the rate of lower-bound readings ('True' responses) decreases under load, while the rate of upper-bound, *exactly* responses increases. This experimental evidence supports the introspective evidence that the doubly-bounded, *exactly* reading of numerals does not have the same underlying basis as the doubly-bounded, *not all* reading for 'some'. Furthermore, the fact that lower-bounded responses drop off under memory load supports the view that this reading is derived pragmatically (Breheny, 2008), although Marty et al. caution that the effect under load may be due to the lower-bound reading for numerals being more inferentially complex (see Buccola & Spector, 2016 for details). On the other hand, *verification* of the lower-bound reading is less complex than for the doubly bounded reading, since checking for four or more red dots only requires finding four red dots, not checking all dots, while checking for exactly four red dots requires checking all dots. So it is unclear that complexity is a factor here.

Turning now to free-choice inferences, illustrated in (8), it was mentioned above that many researchers have argued these are derived using scalar-implicature derivations.

(8) You can have coffee or cake.
~> You can have coffee and you can have cake.

There is less experimental research on free choice than other types of pragmatic effect but one notable result comes from Chemla & Bott (2014). This study present participants with items such as (17) in contexts where it would be false if participants derive the free-choice inference. If they do not derive this inference, they would judge it true:

17. Elsie the engineer is allowed to save a kangaroo or a fork.

Chemla & Bott (2014) report that, unlike the case with scalar implicatures, participants are not delayed in their 'false' judgements compared to their 'true' judgements. Moreover, Chemla & Bott also included unambiguous control items (that have the same truth value irrespective of whether the free choice inference is included). They found an interaction between truth-value and condition such that 'false' judgements were delayed relative to 'true' for the controls, but not the critical judgements. The implications are that, as with numerals, free choice inferences do not have the same underlying basis as other scalar implicatures. This could be because free-choice inferences follow from the conventional meaning of disjunction under certain operators. That is, they are not pragmatic after all.

Alternatively, it could be that free choice inferences are pragmatic but their derivation is not of the same nature as standard Gricean derivation for scalars. Chemla (2010) explores one kind of derivation of free choice effects that make them more like I/R implicatures discussed in section 2.1. By contrast, van Tiel & Schaeken (2016) explore the idea that free-choice inferences are derived as scalars but have a lower cost than typical straight scalars, like (1), because their alternatives are easier to retrieve (being a sub-part of the asserted sentence). van Tiel & Schaeken compared decision times on straight scalar items involving 'some', free choice inferences and also examples of conditional perfection (illustrated in (7) in section 2.1) and clefted sentences. They find that only 'some' items show a delayed response when participants include the respective inference. van Tiel & Schaeken argue that if their three pragmatic inferences were in fact scalar implicatures, what sets them apart from scalars involving 'some' is that all three have easier-to-retrieve alternatives. However, an alternative explanation could be that none of these three pragmatic inferences are

scalar implicatures. This would be consistent with the SAT evidence discussed above (see Bott et al., 2012) that the *not all* inference with 'some' emerges more slowly than the same inference with 'only some' even though the set of alternatives required to interpret both are identical.

To sum up this section on the language-Gricean interface, experimental research has fed into the many debates as to the status of a wide variety of implications. Even relatively simple survey data and reaction-time studies, when designed in a way that is informed by theory, can play a key role in helping theorists to evaluate different accounts of different linguistic constructions.

4. How a Gricean system might integrate linguistic and non-linguistic functions into utterance interpretation

The focus of the experimental research reviewed in the last section has largely been on questions of status – for a given implication, should it be regarded as default, linguistic or Gricean?; is it a scalar implicature or not? In this section, we focus on a different set of questions that take for granted the existence of Gricean pragmatic inferences, including scalar implicatures. These questions have to do with the architecture of a system that combines linguistic information with information about the speaker's goals, beliefs etc., about previous shared information, long-term encyclopaedic memory, and so forth. None of these questions are specific to scalar phenomena but several studies that are relevant to these architectural questions have used scalar phenomena as their subject matter.

4.1. Literal First or Free-for-All?

When it comes to proposals about how linguistic computations and Gricean inferences may be actually integrated in cognition, Grice himself has been widely seen as committing to a 'literal first' view in 'Logic and Conversation' (Grice, 1975), although Searle (1979) is perhaps more definitive on this issue (see Gibbs, 1994; Cacciari & Glucksberg, 1994). In the domain of scalars, this could translate into the idea that the literal unenriched meaning is first accessed prior to an implicature being derived. However, it is by no means necessary that Grice's account of implicature derivation commits us to the temporal priority of the literal in actual utterance processing. This is a point made in Recanati (1995) and elsewhere (see Neale, 1992; Geurts, 2010; Breheny, 2011). It happens that there is much evidence against a literal-first view from research beyond the domain of scalars. In particular, in the domain of metaphor and figurative language processing, there is clear evidence that metaphoric meanings can emerge in the same time course as literal (McElree & Nordlie, 1999) and that they are generated automatically in language processing (Glucksberg et al., 1982).

4.1.1. **The integration of a Gricean System in On-line Utterance Processing**.

One motivation to set ourselves against a literal-first hypothesis comes from the insight that many dimensions of under-specification are ultimately resolved by recourse to the speaker's intentions. That is, when sentences contain lexical or structural ambiguity, or require referential domain restriction by context, as they typically do, one could assume that the ultimate arbiter of how to resolve this ambiguity is by inferring what sense the speaker intended (Wilson & Sperber, 1981). When this assumption is paired with the apparent fact that contextual information is integrated into language processing decisions in the same time-course as linguistically derived information (Altmann & Steedman, 1988; Tanehaus et al., 1995), then it would follow that a Gricean system for inferring speaker meaning should be intimately involved in all levels of language processing, not just the derivation of implicatures. Could this be correct? One set of studies looking at contrastive inferences in reference assignment point to a positive answer to this question. In Sedivy et al. (1999) participants heard, 'Click on the tall glass' when a visual display included a tall glass and another tall object as well as two other objects. When one of those two objects was a shorter glass, participants looked more to the tall glass prior to the onset of the word 'glass', compared to a condition where both other objects were irrelevant distractors. This result shows that participants are exploiting an inference that people typically use modifying adjectives in referential phrases to serve a contrastive function. This inference is arguably Gricean in origin (see Sedivy, 2003 for discussion). Grodner & Sedivy (2007) explore the idea that, as a Gricean inference, the contrast effect ought to be diminished if the participant believes the speaker to have some problems that affect their ability to communicate efficiently. They contrast an 'unreliable speaker' condition with a control condition where the speaker is assumed to have no such problems and find, as a Gricean approach would predict, that the anticipation effect disappears in the unreliable speaker condition. Similarly, Hanna et al. (2003) replicate Sedivy's anticipation effect while participants hear the modifier, 'empty' in a condition where both speaker and hearer share knowledge of the visual display that the participant sees. However in a speaker-ignorant condition, the display is changed without the speaker knowing. In that case, again, the anticipation effect goes away.

These results and others (e.g. Davies & Katsos, 2013; Pogue et al., 2015; Rubio-Fernandez, 2016) demonstrate that comprehenders are able to exploit Gricean inferences during very early on-line processing of inherently underspecified or ambiguous referential expressions. Moreover, exploitation of Gricean inference in language processing is not simply default but influenced by information about the speaker's abilities and also their mental state (see also Arnold et al., 2007;

Heller et al., 2008; Heller et al., 2012). This strongly suggests that the Gricean inferential system is well-integrated into the processing of linguistic stimuli.

More generally, we can briefly consider research on the use of common-ground or perspective taking as further evidence that systems for inferring speakers' intentions are integrated into decisions about definite reference. In one commonly employed perspective-taking task, the participant-hearer and a speaker are jointly aware that the participant sees objects that the speaker does not, while they discuss objects that are commonly visible. As the speaker does not know what kind of objects these private objects may be, she cannot refer to them using phrases like 'the small train'. Thus, participants ought to ignore these privately viewable objects when processing 'the small train' even if one of them is a good fit for that description. Although there is ample evidence that participants find it difficult to entirely ignore the private competitor object (Keysar at al., 2000; Hanna et al., 2003; see Brown-Schmidt & Hanna, 2011), there is also ample evidence that participants' early anticipation of the referent is influenced by what they know about what could be the speaker's intended referent (Hanna et al., 2003; Heller et al., 2008; Brown-Schmidt et al., 2008; Ferguson & Breheny, 2011, among others). That participants' perspective-taking is not perfect is accounted for in terms of constraint-based models of utterance processing (Macdonald et al., 1994; Hanna & Brown-Schmidt, 2011) according to which mental-state information acts as one of many competing constraints in automatic language processing.

4.1.2. Literal First in scalar processing?

Turning now to research on how scalar phenomena are processed on-line, we find a set of apparently mixed results. In the following discussion, one aim is to sort through studies that utilise time-sensitive methods with scalars to determine which might provide insights into how pragmatic scalar inferences are accessed and integrated on-line. To date, this broader, 'How?'-question has been approached via a narrower question concerning the time course of access and integration of scalar inferences. This research is at least relevant to the 'literal first' view of the relation between linguistic and Gricean inferences.

Recall that section 3.2 discussed studies that indicate some kind of delay or cost for scalars. This may lead researchers to conclude that, unlike other pragmatic or context effects, scalars are somehow special in incurring delays. But we should be cautious here. Many of the studies that give rise to these results are not designed to control for factors that may lead to a delayed response but are not relevant to the time course question. For instance, when 'some' is understood as *some and not all* it denotes a function that is in many ways more complex than when it is understood literally. Representing *some and not all of the consultants had a meeting* is more likely to involve

representing both the consultants that had a meeting and those that did not; whereas the literal proposition (lacking the *not all* upper bound) is less likely to. This richer representation of the speaker meaning may contribute to a delayed response in reading time studies (e.g. Breheny et al., 2006) or verification tasks (Bott & Noveck, 2004). Thus, longer latencies in these studies do not clearly support a literal-first view.

Verification studies that compare items with 'some' and 'only some' are potentially more informative on the time course questions since the extra complexity of the upper-bounding *not all* implication in the pragmatically enriched meaning is mirrored in the 'only some' condition. Bott et al. (2012)'s SAT study found a delay in the 'some' case compared to 'only some'. As suggested in section 3.2, this delay could be due to the richer representation for the pragmatic enrichment, involving the speaker's reasons for choosing the weaker alternative. Still, even if the richer epistemic context for the pragmatic condition is responsible for this delay, we need to consider how it bears on our evaluation of the literal-first and other hypotheses about the time course of scalars. It could be that, in virtue of being more complex, a representation of the full meaning of the stimulus takes longer to emerge; or the verification decision task is delayed due to the extra memory demands implied by the more complex representation. Neither of these states of affairs imply that a representation of the literal part of the overall meaning emerges prior to the pragmatic enrichment. Thus, we should look beyond these results when considering the time course question. And so we turn now to studies using visual-world eye-tracking and ERP methods that are potentially more informative on the time course question.

A key methodological contribution on the time course question was made by Huang & Snedeker (2009), who look at embedded scalar enrichments (*EE* from section 2.1).[1] These studies set out to compare the time course of access to and integration of a pragmatic interpretation of noun phrases containing 'some' with an interpretation of items containing 'all', and also with numerical determiners 'two' and 'three'. At least the 'all' condition involves items that do not involve scalar implicature. Participants are asked to point to 'the girl with some of the ...' or 'the girl with all of the...' in a visual display containing two girls. One girl has all of one set of three items (e.g. soccer balls) and one has some but not all of a different set of items – two of four socks. The residue of this

[1] According to the taxonomy in section 2.1, when participants understand, 'Click on the girl with some of the socks' to mean *click on the girl with some and not all of the socks* this cannot be a Straight Scalar but it could be categorised as EE. As mentioned in that section, there are different theoretical views concerning the status of different scalar phenomena. However, it is interesting that neither the broadly Gricean camp (e.g. Geurts & van Tiel, 2013; Bergen et al., 2016) nor the Grammatical camp see EE as involving the same kind of derivation as the standard Gricean derivation of straight scalars. Irrespective of the ways in which some of the scalar enrichments may be derived within a Gricean inferential framework, this literature on the time course of Gricean enrichments takes for granted that these are broadly Gricean, and the discussion proceeds from there. This is the stance adopted in this review discussion also.

partitioned set of socks is with one of two male characters also in the display. Participants preview the visual display prior to the instruction and thus have the opportunity to set up representations of the sets of objects prior to the verbal stimulus. This design therefore eliminates one of the extraneous costs associated with the complexity of the pragmatic interpretation vis-à-vis the semantic, since the relevant representations are available prior to the linguistic input. That is, prior to the instruction, participants not only represent the socks in possession of one of the girls but also the rest of the socks, in the possession of one of the boys. Another key aspect of this design for the 'some' items lies in the fact that if participants process the instruction according to the literal meaning, the referential phrase is ambiguous up until the point of disambiguation in 'socks'. Hence we would expect no bias to form until then. However, if the pragmatic inference can be exploited in the reference assignment process immediately when it becomes available (from the offset of 'some') then bias should form toward the target at the same rate as in the 'all' condition, which is unambiguous from the offset of the determiner.

Huang and Snedeker report that participants' visual bias to the target in the 'some' condition is significantly delayed relative to the 'all' condition. In particular, prior to the point of disambiguation of the noun, participants bias to the target in the 'all' condition is significantly above chance but not in the 'some' condition. In follow up work, Huang & Snedeker (2011) establish that participants can exploit the pragmatic inference prior to the later disambiguation point in the noun, but there is a considerable delay, compared to the 'all' condition. While Huang & Snedeker stop short of endorsing a full, 'literal first' view, they argue that their results are consistent with the idea that the effects of pragmatic inferences must trigger off lexical decoding processes; and this would give some priority to linguistically driven inferences over pragmatic.

In contrast to Huang & Snedeker (2009; 2011), a number of similarly constructed visual-world studies show little or no delay between the 'some' and 'all' condition. Grodner et al. (2010) utilise a very similar design to Huang & Snedeker, involving EE, and find no delay. Breheny et al. (2013a) utilise straight scalars and find no delay of 'some' items relative to 'all'. The reasons for such a disparity in results have yet to be fully clarified. Two kinds of proposal have been made about this disparity, reviewed in Degen & Tanenhaus (2016; see also Grodner et al., 2010).

One difference between the studies finding no delay and those that report a delay relates to the presence of number items. In Huang & Snedeker's experiments, there are two ways of referring to each of the target items (using 'some'/'all' or 'two'/'three'), while in Grodner et al. (2010) and Breheny et al. (2013a) there is only one. It could be that, in the course of experimental sessions, participants come to pre-code targets on the visual display in terms of the quantifiers, 'all' and 'some', bypassing processes that feed off both compositional semantic knowledge and Gricean

inferences. This would explain why bias to targets does not differ between the 'some' and 'all' condition.

An alternative line of explanation for the disparity (see Grodner et al., 2010) holds that the presence of items like 'click on the girl with two socks' undermines the basis for exploiting a Gricean inference in the items that involve 'some' instead of 'two'. Although the reasons are not entirely clear why this would be the case, it could be down to the fact that 'two' is a better cue to the target, since number in the subitizable range are easier to identify visually.

Degen & Tanenhaus (2016) explore these two hypotheses in two visual world studies. In the first, they replicated the results of Grodner et al. (2010) in a study without number items but with various decoy trials that lower the possibility of pre-coding. A second study contained also number items, with the number in question being either in the subitizable range (as used in Huang & Snedeker's studies) or larger, non-subitizable numbers. Degen & Tanenhaus found that bias to target forms faster in 'all' trials than 'some' when the number of items was large but not when the number was small (in the subitizable range). It is interesting that Degen & Tanenhaus' results show that big-set 'all' trials show faster bias formation than small-set 'some', replicating Huang & Snedeker's findings. It seems clear that set sizes in these visual display plays a role in bias formation that has yet to be fully clarified (but see, Chao & Breheny, *under review*).

Overall, these results are certainly problematic for the literal-first view. In Degen & Tanenhaus' without-numbers study, there is less chance of pre-coding, yet Grodner et al.'s results are replicated. In the with-numbers study, comparing targets with small numbers, there is no difference between 'some' and 'all' conditions. So, the pragmatic condition is not necessarily slower than the non-pragmatic condition, even in the presence of number trials. On the other hand, an open question remains about what factors determine results when there is a difference. The expected delay when set size is in the subitizable range did not materialise. In fact, this is the condition where no delay was found.

4.1.3. ERP studies on the time course of scalar implicature.

Like eye-tracking methods, Event-Related Potential (ERP) techniques have been widely employed in psycholinguistics due to the fact that fine-grained resolution can be achieved in terms of the time course of effects (see van Berkum, 2004). ERP studies looking at scalar phenomena have mostly adopted the verification procedure of Bott & Noveck (2004). Nieuwland et al. (2010) used sentences like, 'Some people have lungs/pets....'. In previous research on language processing, anomalous or unexpected stimuli tend to elicit an EEG response characterised by heightened negativity of wave forms centred around 400ms from the onset of the stimulus (an N400 effect). Nieuwland and

colleagues reasoned that if there were rapid access to the scalar implicature (*not all people have lungs/pets*) one might expect an N400 effect when participants encounter the object noun phrase, 'lungs' compared to 'pets'. They test a group of participants and found no overall N400 effect. However, these participants also undertook an Autism Quotient survey (Baron-Cohen et al., 2001) which measures something like sociability across a range of domains, including communication. Nieuwland et al. report that when participants are grouped by sociability according to this index, the high-sociability group do show an N400 effect.

Hunt et al. (2013) report an ERP study with a similar design except that the stimuli were evaluated relative to a visual context. They find that for an item like 'The student cut some of the steaks in this story', there was an expected large N400 effect when none of the steaks were cut (semantically False condition) compared to True controls. In a pragmatically False/semantically True condition (when all steaks were cut) there was an intermediate response. Similar effects are reported in Spychalska et al. (2016).

These ERP studies are informative that scalar inferences are integrated on-line. However, they do not speak directly to question about the early time course of scalar implicature integration because the ERP response measured is triggered by a word that appears some time after the onset of 'some'. In fact, the delay is longer than normal due to the use of one-word-at-a-time presentation procedure necessary for ERP. Moreover, N400 responses are known indicators of surprise or anomaly and do not necessarily indicate the time course of access to linguistic vs. pragmatic inferences. One study that attempts to probe ERP responses on the determiner word itself is reported in Politzer-Ahles et al. (2013). In a verification task, Politzer-Ahles et al. report that when 'some' occurs in an *all* context (i.e. when the sentence would be false under its *some and not all* reading), they found an effect similar to when 'all' appears in a falsifying visual context in an early 200ms-500ms time window. The authors argue that this effect is not like the standard N400 effect. Moreover, unlike the semantically anomalous 'all' items, anomalous 'some' items also gave rise to an effect in a later, 500-1000ms, region. The authors suggest this may reflect reanalysis of 'some', undoing the pragmatic enrichment. Thus there is some ERP evidence for rapid response to pragmatic 'some' in a comparable time course to 'all'. Moreover, the response to inconsistent 'some' has a distinct signature, consistent with the diverse source of the relevant inference – linguistic vs. pragmatic.

To sum up this discussion of time-sensitive research on processing scalar implicatures, the bulk of relevant evidence points toward a system in which inferences derivable from linguistic semantic knowledge and from Gricean pragmatic knowledge can become available in the same time course. There are studies, notably Huang & Snedeker (2009; 2011) that run contrary to this

conclusion but subsequent visual-world research has revealed that factors such as set size impact on the measured response.

4.2. Beyond Time course Studies

The time course studies reviewed in section 4.1 are most relevant to whether language processing is literal first. They shed some light on the broader picture of how scalar implicatures come to be accessed and integrated in on-line processing. However, they leave a lot of questions open. As discussed in Breheny et al., (2013b), if scalar implicature processing is more or less automatic, then results reviewed above raise interesting questions about how this could be. Typically, language processes are conceived of as selecting among choices pre-determined by the rules of the language and lexical information activated by the input stimuli. An ambiguous word or string of words typically leaves a relatively small number of parsing options to choose among, in terms of lexical meaning or syntactic structure. But when it comes to implicatures, particularly ad hoc implicatures (studied in Breheny et al., 2013b), there is no clear sense that any word or string of words should come with a scalar implicature, an R-implicature, an M-implicature etc. On the face of it, it is difficult to conceive of how language processing is set up to continuously decide if any of a set of pragmatic effects should be integrated as linguistic input is incrementally processed. When it comes to scalar implicatures, the decision must involve alternatives and, at least in the case of ad hoc scalars, these would not come precompiled with the lexical input. So, the question is, how would automatic language processes determine when a scalar implicature should be integrated and what it should be?

The beginnings of an answer to this question comes from research into on-line computation of the source of relevance, or 'Question Under Discussion' (QUD). Taking a lead from suggestions in Sperber & Wilson (1986), as well as van Kuppevelt (1995) and Roberts (1996/2012), a series of recent studies provides some evidence that incremental processes anticipate not only the content of the utterance (its semantic and pragmatic interpretation) but also the source of relevance for the utterance, typically described in terms of QUDs (Tian et al., 2010; Clifton & Frazier, 2012; Clifton & Frazier, 2016; Tian et al., 2016; Kehler & Rohde, 2016 – see Breheny, 2018 for discussion).

To illustrate the connection between scalars and QUD, consider an utterance of (18) in a situation where the indicated implication is communicated:

18. Albie ate the salad.

    ~> Albie did not eat the steak or dessert.

The relevant alternatives here (*Albie ate the steak; Albie ate dessert*) could be culled from the question, *which of the salad, steak or dessert did Albie eat?*, since the latter determines a logical space in which Albie eats just steak, just salad, just dessert, steak and salad, salad and dessert, etc. (see Hirschberg, 1991). Imagine that anticipatory incremental processing is set up to compute likely QUDs given linguistic input and context. Then, depending on the prosody of the sentence (see Sperber & Wilson, 1986; Breheny, 2018) input up to the verb in (18) would make *Albie ate what* a more likely relevant question. If, in addition, a small number of edible items are salient in the context (see Altmann & Kamide, 1999) then these become candidates for what is eaten, thereby constituting the alternatives for scalar inference.

Recent work shows a positive impact of prosodic stress, or focus, on rates of scalar implicature (Chevalier et al., 2008; de Marneffe & Tonhauser, 2015; Cummins & Rohde, 2015). Focus constituents have long been thought to reflect the underlying relevant question/QUD in the typical case (Sperber & Wilson, 1986; Roberts, 1996/2012; Breheny, 1996; 1998 – see Cummins & Rohde, 2015 for discussion). So this off-line data provides indirect support for the idea that QUDs play a role in determining which scalars may be derived from an utterance.

To date, little work has been completed that examines on-line computation of QUDs and scalars together, but the prospects are promising that for at least one set of Gricean inferences, we will arrive at a better understanding of the factors that impact on their availability in incremental processing.

5. Conclusions – Back to the origins of a Gricean System

In the last section, I speculated that the question of how scalar implicatures are integrated as part of incremental processing might be answered via a theory of incremental anticipation of a specific feature of context – the source of relevance, or QUD. As mentioned, there is research beyond the domain of scalars suggesting that QUD anticipation may be part of incremental language processing and this may bear on more explicit aspects of the message such as reference assignment (Kehler & Rohde, 2015). In closing this review, I will discuss how evidence from disparate areas of research can be brought together to build a coherent picture of a human Gricean system.

A most impressive difference between human infants, on the one hand, and chimpanzees or other great apes, on the other, lies not in their ability to represent goals or intentions, knowledge states and the like (Call & Tomasello, 2008), but in their response to communicative stimuli. One stark illustration of this difference comes with the so-called object choice task – where the participant is aware the target is in one of two locations and the experimenter communicatively points to the correct location. Behne et al. (2005) report that 14 month-old infants interpret the

signal correctly, while chimpanzees respond at chance, even though they notice the stimulus (Call & Tomasello, 2005) and they have the same goal to find the target. Although different accounts of these differences have been given (Tomasello et al., 2005; Gergely & Csibra, 2005) they have in common the idea that humans expect a communicative signal to be relevant in a shared context. That communicative signals engender expectations of relevance has been demonstrated in pre-verbal infants from at least 10 months of age (Yoon et al., 2009; Kovacs et al., in press), while toddlers around one year of age have been shown to rely on expectations of relevance in a shared context to interpret communicative gestures (Liebal et al., 2010; Southgate et al., 2009). Moving to later stages of language/communicative development, it has been shown that children given a sentence verification task need to be provided with a context of plausible denial (Crain et al., 1996). This is such that, if the target sentence is *S* then the utterance context ought to be such that whether *S* is true is a plausible question under discussion.

Given recent adult processing research and also child developmental research, a reasonable hypothesis about a human Gricean system is that it develops to process communicative stimuli, including linguistic utterances, in a way that determines the relevance of that stimulus in a shared context. Just how we should define relevance, be it in terms of a formal structure involving questions and sub-questions (Roberts, 1996/2012; Ginzburg, 2012), or in terms of utterance utility (Sperber & Wilson, 1986; van Rooy, 2003; Csibra & Gergely, 2011), or in terms of impact on the likelihood of a topic proposition (Carnap, 1950; Russell, 2012) remains to be determined. These analytical alternatives have descriptively similar coverage, including when it comes to scalars.

Assuming a relevance-oriented hypothesis about the cognitive system underpinning utterance processes, the aim of future research would be to investigate what factors impact on the determination of the source of relevance for an utterance, on the content of the utterance itself and the interplay between these factors.

References

Altmann & Steedman (1988). Interaction with context during human sentence processing. *Cognition*, 30, 191-238.

Altmann, G.T.M., and Kamide, Y . (1999) Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition* , 73(3), 247-264.

Antoniou, K., Cummins, C., & Katsos, N. (2016). Why only some adults reject under-informative utterances. *Journal of Pragmatics*, 99, 78-95.

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

Barker, C. (2010). Free choice permission as resource-sensitive reasoning. *Semantics and Pragmatics*, 3(10), 1–38.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism spectrum quotient (AQ): Evidence from Asperger syndrome/high functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, 31, 5–17.

Benz, Anton & Robert van Rooij. (2007). Optimal assertions and what they implicate. *Topoi* 26(1). 63–78.

Bergen, L. & Grodner, D. (2012). Speaker Knowledge Influences the Comprehension of Pragmatic Inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition*.

Bergen, L. Levy, R. and Goodman, N. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*.

Van Berkum, J. J. A. (2004). Sentence comprehension in a wider discourse: Can we use ERPs to keep track of things? In M. Carreiras, Jr. & C. Clifton, Jr. (Eds.), *The on-line study of sentence comprehension: Eyetracking, ERPs and beyond* (pp. 229–270). New York: Psychology Press.

Bott, L., & Noveck, I. A. (2004). Some utterances are underinformative: The onset and time course of scalar inferences. *Journal of Memory and Language*, 51(3), 437–457. doi:10.1016/j.jml.2004.05.006

Bott, L., Bailey, T. M., & Grodner, D. (2012). Distinguishing speed from accuracy in scalar implicatures. *Journal of Memory and Language*, 66(1), 123–142.

Breheny, R. (1996). Pro-active focus. In *UCL Working Papers in Linguistics*, 8,

Breheny, R. (1998). Interface economy and focus. In Villy Rouchota & Andreas H. Jucker (eds), *Current Issues in Relevance Theory*. John Benjamins, pp105-140.

Breheny, R. (2008). A new look at the semantics and pragmatics of numerically quantified noun phrases. *Journal of Semantics*, 25 (2), 93.

Breheny, R. (2011). Experimentation-Based Pragmatics. In: Bublitz, W and Norrick, N, (eds.) *Handbook of Pragmatics: Volume 1 Foundations of Pragmatics*. Mouton – de Gruyter.

Breheny, R. (2018). Language processing, relevance and questions. To appear in K. Scott, B. Clark & R. Carston (eds), *Relevance: Pragmatics and Interpretation*. CUP.

Breheny, R., Katsos, N., Williams, J., 2006. Are generalised scalar implicatures generated by default? An on-line investigation into the role of context in generating pragmatic inferences. *Cognition* 100 (3), 434--463.

Breheny, R., Ferguson, H.J., & Katsos, N. (2013a). Investigating the timecourse of accessing conversational implicatures during incremental sentence interpretation. *Language And Cognitive Processes*, 28, 443-467. doi:10.1080/01690965.2011.649040

Breheny, R., Ferguson, H.J., Katsos, N., (2013b). Taking the epistemic step: toward a model of on-line access to conversational implicatures. *Cognition* 126 (3), 423--440.

Breheny, R., Klinedinst, N., Romoli, J & Sudo, Y. (forthcoming) The symmetry problem: current theories and prospects. *Natural Language Semantics*.

Brown, P., & Levinson, S. (1987). *Politeness: Some language universals in language use*. Cambridge: Cambridge University Press.

Brown-Schmidt, S., Gunlogson, C., & Tanenhaus, M. K. (2008). Addressees distinguish shared from private information when interpreting questions during interactive conversation. *Cognition*, 107, 1122–1134.

Brown-Schmidt, S., & Hanna, J. E. (2011). Talking in another person's shoes: incremental perspective-taking in language processing. *Dialog and Discourse*, 2, 11–33.

Buccola, B. & Spector, B. (2016). Modified numerals and maximality. *Linguistics and Philosophy* 39: 151. doi:10.1007/s10988-016-9187-2.

Cacciari, C.,& Glucksberg, S. (1994). Understanding figurative language. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 447-477). New York: Academic Press.

Call, J., & Tomasello, M. (2005). What do chimpanzees know about seeing revisited: an explanation of the third kind. In N. Eilan, C. Hoerl, T. McCormack, & J. Roessler (Eds.), *Issues in joint attention* (pp. 45–64). Oxford: Oxford University Press.

Call, J. & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Science*, 12, 187-192.

Carnap, R., (1950). *Logical Foundations of Probability*. University of Chicago Press, Chicago.

Chao, S. & Breheny, R. (under review). What would a compostional listener do? – Another look at the timecourse of scalar implicatures.

Chemla, E. (2010) Similarity: Towards a unified account of scalar implicatures, free choice permission and presupposition projection.

Chemla, E., & Bott, L. (2014). Processing inferences at the semantics/pragmatics frontier: disjunctions and free choice. *Cognition*, 130(3), 380–396.

Chevallier, C., Noveck, I. A., Nazir, T., Bott, L., Lanzetti, V., and Sperber, D. (2008). Making disjunctions exclusive. *Q. J. Exp. Psychol*. 61, 1741–1760. doi: 10.1080/17470210701712960.

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

Chierchia, G., Fox, D., & Spector, B. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In P. Portner, C. Maienborn, & K. von Heusinger (Eds.), *An international handbook of natural language meaning* (pp. 2297–2332). Berlin: Mouton de Gruyter.

Clifton, C., & Frazier, L. (2012). Discourse integration guided by the "question under discussion." *Cognitive Psychology*.

Clifton, C., & Frazier, L. (2016). Accommodation to an unlikely episodic state. *Journal of Memory and Language*, 86, 20–34. doi:10.1016/j.jml.2015.10.004.

Crain, S., Thornton, R. Boster, C. Conway, L. Lillo-Martin, D. and Woodams, E. (1996). Quantification without qualification. *Language Acquisition* 5, 83-153.

Csibra G. (2010). Recognizing communicative intentions in infancy. *Mind & Language*. 25(2):141-68.

Csibra G. & Gergely G. (2011). Natural pedagogy as evolutionary adaptation. *Philosophical Transactions of the Royal Society B*.;366:1149-57.

Cummins, C. & Rohde, H. (2015). Evoking context with contrastive stress: Effects on pragmatic enrichment. *Frontiers in Psychology*, Special issue on Context in communication: A cognitive view, 6(1779), 1-11.

Davies, C and Katsos, N (2013) Are speakers and listeners 'only moderately Gricean'? An empirical response to Engelhardt et al. (2006). *Journal of Pragmatics*, 49 (1). 78 - 106.

De Neys W., & Schaeken W. (2007). When people are more logical under cognitive load: Dual task impact on scalar implicature. *Experimental Psychology*, 54(2), 128–133.

Degen, J. (2015). Investigating the distribution of "some" (but not "all") implicatures using corpora and webbased methods. *Semantics and Pragmatics*, 8(11), 1–55.

Degen, J., & Tanenhaus, M. K. (2015). Processing scalar implicature: A constraintbased approach. *Cognitive Science*, 39(4), 667–710.

Degen, J., & Tanenhaus, M. K. (2016). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, 40(1), 172–201.

Dieussaert, K., Verkerk, S., Gillard, E., Schaeken, W., (2011). Some effort for some: further evidence that scalar implicatures are effortful. *Q. J. Exp. Psychol*. 64 (12), 2352--2367.

Duffy, S. Morris R. & Rayner K. (1988). Lexical ambiguity and fixation times in reading, *Journal of Memory and Language* 27 429–446.

Feeney, A., Scrafton, S., Duckworth, A., & Handley, S. J. (2004). The story of some: Everyday pragmatic inference by children and adults. *Canadian Journal of Experimental Psychology*, 58(2), 121–132. doi:10.1037/h0085792.

Ferguson, H. J., Breheny, R. E. T. (2011). Listeners' eyes reveal spontaneous sensitivity to others' perspectives. *Journal of Experimental Social Psychology*, 48 (1), 257-263.

Fox, D. (2007). Free choice and the theory of scalar implicatures. In U. Sauerland & P. Stateva (Eds.), *Presupposition and implicature in compositional semantics* (pp. 71–120). Houndmills: Palgrave Macmillan.

Fox, D., & Katzir, R. (2011). On the characterization of alternatives. *Natural Language Semantics*, 19(1), 87–107.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336, 998.

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

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

Geurts, B. (2005). Entertaining alternatives: Disjunctions as modals. *Natural Language Semantics*, 13(4),383–410. doi:10.1007/s11050-005-2052-4.

Geurts, B. (2006). Take 'five': The meaning and use of a number word. In S. Vogeleer & L. Tasmowski (Eds.), *Non-definiteness and plurality* (pp. 311–329). Amsterdam and Philadelphia, PA: John Benjamins.

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

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

Gergely, G. and Csibra, G. 2003: Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Science*, 7(7), 287–292.

Gergely G, Csibra G. (2005). A few reasons why we don't share Tomasello et al.s intuitions about sharing. *Behavioral and Brain Sciences*. 28(5):701-2.

Gibbs, R. w., Jr. (1994). Figurative thought and figurative language. In M. A. Gernsbacher (Ed.), *Handbook ofpsycholinguistics* (pp. 411446). New York: Academic Press.

Ginzburg, J. (2012). *The interactive stance: meaning for conversation*. CSLI: Center. Oxford University Press.

Glucksberg, S., Gildea, P.. & Bookin. M. B. (1982). On understanding nonliteral speech: Can people ignore metaphors. *Journal of Verbal Learning & Verbal Behavior*, 21, 85-98.

Grice, H.P. (1957). Meaning, *The Philosophical Review*, 66: 377–88..

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

Grice, H.P. (1978). Further Notes on Logic and Conversation, in *Syntax and Semantics: Pragmatics*, v 9, P. Cole (ed.), New York: Academic Press, 183–97.

Grodner, D. J., Klein, N. M., Carbary, K. M., & Tanenhaus, M. K. (2010). “Some,” and possibly all, scalar inferences are not delayed: Evidence for immediate pragmatic enrichment. *Cognition*, 116(1), 42–55. doi:10.1016/j.cognition.2010.03.014.

Grodner, D. J., & Sedivy, J. C. (2007). The effect of speaker-specific information on pragmatic inferences. In E. Gibson & N. Perlmutter (Eds.), *The processing and acquisition of reference*. Cambridge, MA: MIT Press.

Hanna JE, Tanenhaus MK, Trueswell JC. (2003). The effects of common ground and perspective on domains of referential interpretation. *Journal of Memory and Language*. 49:43–61.

Heller D, Grodner D, Tanenhaus MK. (2008). The role of perspective in identifying domains of reference. *Cognition*. 2008;108:831–836.

Heller D., Gorman K. S., Tanenhaus M. K. (2012). To name or to describe: shared knowledge affects referential form. *Top. Cogn. Sci*. 4 290–305. 10.1111/j.1756-8765.2012.01182.x.

Herrmann, E., Call, J., Hernandez-Lloreda, M.V., Hare, B., Tomasello, M. (2007). Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis. *Science.* Vol. 317, Issue 5843, pp. 1360-1366. DOI: 10.1126/science.1146282

Hirschberg, Julia Bell. 1991. *A Theory of Scalar Implicature*. New York: Garland.

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

Horn, L. R.: 1989, *A Natural History of Negation*, University of Chicago Press, Chicago, IL.

Horn, L., R. (1992). The said and the unsaid. *SALT II*, 163-192, Dept. of Linguistics, Ohio State University.

Huang, Y. T., & Snedeker, J. (2009). On-line interpretation of scalar quantifiers: Insight into the semantics-pragmatics interface. *Cognitive Psychology*, 58(3), 376–415. doi:10.1016/j.cogpsych.2008.09.001.

Huang, Y. & Snedeker, J. (2011). ‘Logic & Conversation’ revisited: Evidence for a division between semantic and pragmatic content in real time language comprehension. *Language and Cognitive Processes*, 26, 1161-1172.

Ippolito, M. (2007). On the meaning of some focus-sensitive particles. *Natural Language Semantics*, 15(1): 1-34.

Katzir, R. (2007). Structurally-defined alternatives. *Linguistics and Philosophy*, 30(6), 669–690. doi:10.1007/s10988-008-9029-y.

Kehler, A. & Rohde, H. (2016). Evaluating an Expectation-Driven QUD Model of Discourse Interpretation. *Discourse Processes*.

Kennedy, C. (2015). A “de-Fregean” semantics (and neo-Gricean pragmatics) for modified and unmodified numerals. In: *Semantics and Pragmatics* 8.10, pp. 1–44. doi: 10.3765/sp.8.10.

Klinedinst, N. (2007). Plurals, possibilities, and conjunctive disjunction. *UCL Working Papers in Linguistics*, 19, 261–284.

Kovács ÁM, Téglás E, Gergely G, Csibra G. (In Press). Seeing behind the surface: Communicative demonstration boosts category disambiguation in 12-month-olds. *Developmental Science*.

Kuppevelt, J. van. (1995). Discourse structure, topicality, and questioning. *Journal of Linguistics*, 31, 109-147.

Levinson, S. C. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

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

Lewis, D. (1969). *Convention*: *A Philosophical Study*, Cambridge, MA: Harvard University Press.

Liebal, K.., Behne, T., Carpenter, M., & Tomasello, M. (2009). Infants use shared experience to interpret pointing gestures. *Developmental Science*, 12, 264-71.

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

McElree,B.,& Nordlie,J. (1999). Literal and figurative interpretations are computed in equal time. *Psychon.Bull.Rev*. 6, 486–494.doi:10.3758/BF03210839

Magri, Giorgio. (2009). A theory of individual-level predicates based on blind mandatory scalar implicatures. *Natural Language Semantics* 17(3). 245–297. doi:10.1007/s11050-009-9042-x.

de Marneffe, M.-C., and Tonhauser, J. (2015). "On the role of context and prosody in the generation of scalar implicatures of adjectives," in Presentation at XPRAG.de Workshop "Formal and Experimental Pragmatics: Methodological Issues of a Nascent Liaison," Berlin.

Paul Marty and Emmanuel Chemla (2013). Scalar implicatures: working memory and a comparison with 'only'. *Frontiers in Psychology* 4(403), doi:10.3389/fpsyg.2013.00403.

Paul Marty, Emmanuel Chemla & Benjamin Spector (2013). "Interpreting numerals and scalar items under memory load". *Lingua* 133, pp 152-163, doi:http://dx.doi.org/10.1016/j.lingua.2013.03.006.

Paul Marty, Emmanuel Chemla & Benjamin Spector (2015). "Phantom readings: the case of modified numerals". *Language, Cognition and Neuroscience* 30(4), pp 462-477.

Matsumoto, Y.: 1995, 'The Conversational Condition on Horn Scales', *Linguistics and Philosophy* **18**, 21–60

Merin, A.: 1999, 'Information, Relevance, and Social Decisionmaking', in L. Moss et al. (eds.), *Logic, Language, and Computation*, Vol. 2, pp. 179–221, CSLI publications, Stanford.

Moll, H., & Meltzoff, A. N. (2011). Perspective-taking and its foundation in joint attention. In N. Eilan, H. Lerman, & J. Roessler (Eds.), *Perception, Causation, and Objectivity. Issues in Philosophy and Psychology* (pp. 286-304).Oxford: Oxford University Press.

Neale, S. (1992). Paul Grice and the Philosophy of Language. *Linguistics and Philosophy* 15 (5):509 - 559.

Nieuwland, M.S., Ditman, T., Kuperberg, G.R., (2010). On the incrementality of pragmatic processing: an ERP investigation of informativeness and pragmatic abilities. *J. Mem. Lang*. 63 (3), 324--346.

Noveck, I. A., & Posada, A. (2003). Characterizing the time course of an implicature: An evoked potentials study. *Brain and Language*, 85(2), 203–210. doi:10.1016/s0093-934x(03)00053-1.

Pogue A., Kurumada C., Tanenhaus M. K. (2015). Talker-specific generalization of pragmatic inferences based on under– and over-informative prenominal adjective use. *Front. Psychol*. 6:2035 10.3389/fpsyg.2015.02035.

Politzer-Ahles S, Fiorentino R (2013) The Realization of Scalar Inferences: Context Sensitivity without Processing Cost. *PLoS ONE* 8(5): e63943. doi:10.1371/journal.pone.0063943.

Poppels, T. and Levy, R. (2015). Resolving quantity and informativeness implicature in indefinite reference. *Proceedings of the 2015 Amsterdam Colloquium: The Workshop on Reasoning in Natural Language*, pp. 313–322. Zuidema, Willem and Szymanik, Jakub (ed).

Potts, Christopher; Daniel Lassiter; Roger Levy; and Michael C. Frank. 2015. Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. To appear in *Journal of Semantics*.

Recanati, F. (1995). The Alleged Priority of Literal Intepretation. *Cognitive Science* 19: 207-232.

Roberts, C. (1996). Information structure in discourse: Towards an integrated formal theory of pragmatics. OSU Working Papers in Linguistics, 49: Papers in Semantics. (Reprinted in *Semantics & Pragmatics*, 2012)

Romoli, Jacopo. 2013. A problem for the structural characterization of alternatives. *Snippets* 27. 14–15.

Rubio-Fernandez, P. (2016). How Redundant Are Redundant Color Adjectives? An Efficiency-Based Analysis of Color Overspecification. *Frontiers in Psychology*. 7: 153. doi: 10.3389/fpsyg.2016.00153

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

Sauerland, Uli. 2008. Implicated presuppositions. In Anita Steube (ed.), *The discourse potential of underspecified structures*, 581-600. Berlin: Mouton de Gruyter.

Schulz, Katrin & Robert van Rooij. 2006. Pragmatic meaning and nonmonotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy* 29(2). 205–250. doi:10.1007/s10988-005-3760-4.

Searle, J. (1979). Metaphors. In A. Ortony (Ed.), *Metaphor and thought* (pp. 92-123). Cambridge: Cambridge University Press.

Sedivy, J. (2003). Pragmatic versus Form-based Accounts of Referential Contrast: Evidence for Effects of Informativity Expectations. *Journal of Psycholinguistic Research*, Vol. 32(1).

Sedivy, J., Chambers, C., Tanenhaus, M., & Carlson, G. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, 109-147.

Simons, M. (2001). On the conversational basis of some presuppositions. In R. Hastings, B. Jackson and Z. Zvolensky, eds., *Proceedings of Semantics and Linguistic Theory 11*, Ithaca, NY: CLC Publications: 431–448.

Soames, S. (1982). How Presuppositions are Inherited: A Solution to the Projection Problem, *Linguistic Inquiry* **13**, 483–545.

Southgate, V., Chevallier, C., & Csibra, G. (2009). Sensitivity to communicative relevance tells young children what to imitate. *Developmental Science*, 12, 1013-1019.

Southgate V, Chevallier C, Csibra G. (2010). Seventeen-month-olds appeal to false beliefs to interpret others' referential communication. *Developmental Science*.13(6):907-12.

Sperber, D., Wilson, D., 1986. *Relevance: Communication and Cognition*. Blackwell, Oxford.

Sperber, D., Wilson, D., 1995. Postface to Second Edition. In *Relevance: Communication and Cognition*. Blackwell, Oxford.

Tanenhaus, M., Spivey-Knowlton, M., Eberhard, K., & Sedivy, J. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, 268, 1632-1634.

Tian, Y., Breheny, R., & Ferguson, H. (2010). Why we simulate negated information: A dynamic pragmatic account. The Quarterly Journal of Experimental Psychology, 63(12), 2305–2312.

Tian, Y., Ferguson, H., & Breheny, R. (2016). Processing negation without context – why and when we represent the positive argument. Language, Cognition and Neuroscience. doi:10.1080/23273798.2016.1140214

van Tiel, B. & Schaeken, W. (2016). Processing conversational implicatures: alternatives and counterfactual reasoning. To appear in: *Cognitive Science*. DOI: 10.1111/cogs.12362.

Tiel, B. van, Miltenburg, E. van, Zevakhina, N., Geurts, B. (2016). Scalar diversity. Journal of Semantics, 33 (1), 107-135.

Tomasello, M. (2008). *Origins of Human Communication*. MIT Press.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 28, 675 - 691.

Tomlinson Jr., J. M., Bailey, T. M., & Bott, L. (2013). Possibly all of that and then some: Scalar implicatures are understood in two steps. *Journal of Memory and Language*, 69(1), 18–35. doi:10.1016/j.jml.2013.02.003.

Trinh, T. & Haida, A. (2015). Constraining the derivation of alternatives. *Natural Language Semantics*.

van Rooij, R. (2003). Questioning to resolve decision problems, *Linguistics and Philosophy*, 26: 727-763.

van Rooij, R. & K. Schulz (2007). Only: meaning and implicature. In M. Aloni, A. Butler & P. Dekker (Eds.), *Questions in dynamic semantics* (Current research in the semantics/pragmatics interface, 17) (pp. 193-224). Amsterdam: Elsevier.

Wilson, D. (1975). *Presuppositions and Non-Truth-Conditional Semantics*. Academic Press.

Wilson, D. & D. Sperber (1981). On Grice's Theory of Conversation, in P. Werth (ed.), *Conversation and Discourse*, Croom Helm, London, pp. 155-178.