Chapter 5
Neo-Griceanism and its rivals

The previous chapters assessed the Gricean framework as a theory of implicature generation — of what makes it the case that certain utterances carry implicatures. I argued that the framework is best understood as aiming to provide a normative, speaker-independent notion of conversational implicature, parallel to that of conventional meaning. And I went on to argue that it fails in that aim: Gricean principles cannot be applied without appealing at some level to speaker intentions. I also argued that theories of implicature generation cannot be separated entirely from psychological theories of how implicatures are recovered and that it is hard to see how Grice's account could be integrated with such a theory.

But we should not write off the Gricean framework yet. Although it may not succeed in its original aim, it may still explain some important aspects of conversational implicature. In fact, many linguists have seen Grice's ideas as providing the basis for accounts of implicature recovery. These *neo-Gricean* theorists argue that in interpreting utterances we automatically apply certain general principles or heuristics, which are versions of the Gricean maxims. Where applicable, these principles yield non-literal meanings that become the default, or preferred, interpretations of utterances of the type in question. The same principles, it is assumed, guide speakers in their choice of utterance, ensuring that communication is usually successful (for example, Levinson 2000, p.24). Thus, this view treats the Gricean maxims, not as norms of conversation, but as inferential principles that guide conversational behaviour (Levinson 2000, p.35).

Though neo-Griceanism is primarily a theory of implicature recovery (of how hearers derive implicatures), it can also be seen as offering a theory of implicature generation. Neo-Griceans can say that an utterance possesses a generalized implicature q if the interpretive principles the theory posits would yield q as a preferred interpretation of it. Since neo-Griceans hold that we do typically apply these principles in making and interpreting utterances, this amounts to saying that an utterance implicates q if hearers would typically interpret it as possessing it and speakers typically would expect it to. Understood in this way, the theory would licence normative claims to the effect that a *particular* speaker or hearer has

misunderstood what an utterance implicates. Since the neo-Gricean principles are derived from Grice's maxims, and make many of the same predictions, this approach promises to rescue at least part of the original Gricean project. (Neo-Griceans sometimes speak of the principles they posit 'generating' implicatures — meaning that they yield them as interpretations in the minds of hearers. Although I treat neo-Griceanism primarily as a theory of implicature recovery, I shall occasionally follow this usage in the chapter, especially since, as just explained, the principles can also be thought of as generating implicatures in the constitutive sense.)

This chapter will assess neo-Griceanism, focusing on one prominent version of the approach. The first section outlines the theory, and the second section briefly introduces some rival theories with which I shall contrast it. The third and fourth sections look at the neo-Gricean principles, asking whether they accord with and explain our intuitions about what implicatures utterances possess, and the final section reviews some relevant experimental work.

The literature in this area is often technical and deeply involved with wider issues in theoretical linguistics. It is impossible to do justice to it in a chapter, but I shall focus on some key points and test cases.

1. Neo-Griceanism

Neo-Gricean theories propose simple, formalized versions of the Gricean maxims and rules for their application, with the aim of explaining and predicting patterns of implicature. Key figures in the field are Gerald Gazdar (1979), Laurence Horn (1984, 1989, 2004), and Stephen Levinson (1983, 2000). I shall focus on Levinson's presentation in his 2000 book *Presumptive Meanings*, which synthesizes earlier work in the tradition and is a comprehensive and influential presentation of the neo-Gricean approach (Levinson 2000).

1.1 Utterance-type meaning

Levinson adopts a broadly Gricean approach to communication, distinguishing aspects of meaning that are coded (including what is said and what is conventionally implicated) and aspects that are inferred from conversational

principles, including, but not limited to, conversational implicatures (2000, p.14).[1] Levinson follows Grice in accepting that there are two types of conversational implicature — particularized and generalised. (Levinson uses the abbreviations PCI and GCI.) PCIs hold *because of* specific contextual assumptions that do not hold in all, or even many, cases; whereas GCIs hold universally *unless* there are specific contextual assumptions that cancel them. Take, for example:

(1) Some of the guests got food-poisoning.

Unless the implicature is cancelled (say, by adding 'In fact all of them did') this would always carry the implicature that not all the guests got food poisoning, which is a GCI. However, if uttered in response to the question 'How was the wedding?' it would also carry the implicature 'The wedding went badly', or something similar, which would be a PCI. Levinson suggests that PCIs are the result of applying the maxim of Relevance, where this involves attending to the particular speaker's goals and plans (2000, p.17, p.380 n.4).

Levinson's primary interest is in GCIs, which he regards as the central class of 'presumptive meanings' — default, or preferred, interpretations, which are 'carried by the structure of utterances, given the structure of the language, and not by virtue of the particular contexts of utterance' (2000, p.1). These presumptive meanings, he claims, form a distinct level of meaning, *utterance-type meaning*, which is distinct from both the linguistically coded meaning that a sentence carries in every context (sentence-meaning), and the pragmatically enriched meaning that a sentence carries when uttered by a particular speaker in a particular context (speaker-meaning, or utterance-token meaning). Utterance-type meaning is like

---

[1] Although he makes this distinction, Levinson regards all communication as, fundamentally, an inferential process, in which even coded aspects of meaning are clues to interpretation. He writes:

> From a Gricean perspective, communication involves the inferential recovery of speakers' intentions: it is the recognition by the addressee of the speaker's intention to get the addressee to think such-and-such that essentially constitutes communication. (Levinson 2000, p.29)

sentence-meaning in being context-independent and deeply connected with the structure of language, but unlike it in being cancellable. Utterance-type meaning is like speaker-meaning in resulting from pragmatic enrichment of sentence-meaning, but unlike it in resulting from the application of general principles rather than theorizing about the speaker's intentions (2000, p.22). (Levinson notes that utterance-type meaning is not composed only of GCIs, but includes a variety of other pragmatic phenomena, including presuppositions, conventional implicatures (in Grice's sense), and conventions of use (2000, p.23). Levinson's aim is to defend the existence of a level of utterance-type meaning, in opposition to *reductionists*, who would reduce it either to sentence meaning or (as relevance theorists do) to speaker meaning (2000, p.25).

Levinson suggests that our capacity for GCIs is an evolutionary adaption, which developed in order to compensate for an inefficiency in human communication (2000, pp.27–9). He points out that pre-articulation and comprehension processes in the human brain run three to four times faster than the process of phonological articulation (2000, p.28). Our brains can prepare and process utterances much faster than our vocal systems can articulate them, creating a bottleneck in the human communication system. Evolution has eased this bottleneck, Levinson argues, by designing our comprehension systems to apply certain general pragmatic principles, or heuristics, which are defined over formal features of utterances and are applied by default whenever certain expressions are encountered.[2] These principles yield GCIs — default interpretations that are derived without the need to theorize about the speaker's intentions — and they speed up communication by creating an extra layer of utterance-type meaning,

---

[2] Levinson writes:

> Now, the solution to the encoding bottleneck, I suggest, is just this: let not only the content but also the metalinguistic properties of the utterance (e.g., its form) carry the message. Or, find a way to piggyback meaning on top of the meaning ... by utilizing the form, the structure, and the pattern of choices within the utterance to signal the extra information beyond the meanings of its constituents. (Levinson 2000, p.6)

which enriches the content of utterances in ways that we all understand and expect. This requires, of course, that GCIs are recovered very swiftly, without complex theorizing. Indeed, Levinson suggests that some of the principles are applied on a word-by-word basis and that a quantifier such 'some' will trigger its default interpretation of 'not all' even before the predicate it governs has been processed (2000, p.5, p.259)[3].

Although the inferencing here is default, Levinson stresses that it is *defeasible*; it goes through automatically unless contrary information is available, in which case the inference is cancelled. The process therefore cannot be a deductive one, since deductive inference is not defeasible, and it must involve some form of non-monotonic reasoning. Levinson reviews various types of defeasible inference, including induction, abduction, practical reasoning, and default logics, and argues that the last offers the most promising model for implicature (2000, pp.45–6).

1.2 The three principles

Levinson proposes that GCIs can be accounted for by appeal to three principles, which he calls the Q-principle, the I-principle, and the M-Principle — the first and second derived from Grice's maxim of Quantity and the third from his maxim of Manner. The maxim of Quality plays 'only a background role' in the production of GCIs, and the maxim of Relation or Relevance plays none (it 'has pertinence only to the immediate, ever variable, conversational goals: it generates PCIs, not GCIs') (Levinson 2000, p.74).

The Q-principle can be summarized as 'What isn't said, isn't.' The idea is that speakers make the most informative statement they can, given what they know, and that hearers assume that they do this. This resembles Grice's first submaxim of Quantity ('Make your contribution as informative as is required'). On its own the Q-principle is too vague to be applied automatically, and Levinson explains

---

[3] Levinson writes 'the phrase *some of the boys* can invoke the default assumption "not all of the boys" even before the predicate has been heard' (Levinson 2000, p.259). This is an important claim for him, since he holds that it is just these sorts of rapid default inferences that enable hearers to transform semantic fragments into full-blown propositional representations (Levinson 2000, pp.256–9).

that it can be applied only where there is a salient set of contrasting expressions of different informational strength from which the speaker is assumed to have chosen. In choosing a weaker element from the set, the speaker implicates that a corresponding statement substituting a stronger element is false. So, for example, if a person utters the sentence, 'Some of the students failed', the contrast set *<all, some>* is salient, and the Q-principle produces the default reading 'It is not the case that all the students failed.'

The most important class of Q-implicatures are *scalar implicature*s, which depend on an entailment scale (also known as a Horn scale), in which the stronger elements entail the weaker ones. As examples, Levinson gives the following (with stronger items to the left):

> quantifiers *<all, most, many, some>*,
> connectives *<and, or>*,
> modals *<necessarily*, *possibly>*, *<must, should, may>*,
> adverbs *<always, often, sometimes>*,
> degree adjectives *<hot, warm>*
> verbs *<know, believe>*, *<love, like>*
> (Levinson 2000, p.79).

In making an utterance using an expression to the right of one of these scales, one Q-implicates that a corresponding utterance substituting an expression to the left either is false, or might be for all one knows.[4] Thus 'Some of the students passed' Q-implicates that not all the students passed'; 'I told Jack or Annie' Q-implicates

---

[4] There is debate about the strength of the epistemic commitment involved in Q-implicatures. Does the speaker implicate that they know that the stronger claim is false, or that they believe it is, or simply that it may be false for all they know? The issues are complex and I shall not address them here. (For discussion and references, see Levinson 2000, pp.77–9.) In any case, as Atlas notes, the speaker's attitude is independent of the *content* implicated. An utterance of 'Some F are G' implicates that not all F are G, and invites the hearers to believe that not all F are G, regardless of what attitude the speaker is understood to take towards that claim (Atlas 1993, discussed in Levinson 2000, p.78, p.387, n.10).

that the speaker didn't tell both Jack and Annie; 'You may smoke' Q-implicates that it is not the case that the hearer is obliged to smoke, and so on.

The other main type of Q-implicature is *clausal implicature*, which can arise when a sentence contains an embedded clause. By choosing an expression that does not entail the truth of the embedded clause instead of one that does, the speaker implicates that they do not know whether or not the embedded clause is true. For example, if I say 'John believes there is life on Mars' (rather than 'John *knows* there is life on Mars') I Q-implicate that I myself do not know whether or not there is life on Mars (Levinson's example; 2000, p.76).[5]

The second principle is the I-principle, which Levinson summarizes as 'What is simply described is stereotypically exemplified'. The idea here is that typical ('unmarked') expressions implicate that the thing described is itself typical, prompting hearers to fill in the details according to the appropriate stereotype. Levinson notes that this is related to Grice's second sub-maxim of Quantity: 'Do not make your contribution more informative than is required'. Speakers need not spell out details that hearers will fill in automatically ('one need not say what can be taken for granted') (2000, p.37).

The I-principle is a powerful one, which underpins a variety of linguistic phenomena, including *generality narrowing*, where a general expression is interpreted in a more specific sense ('secretary' is understood as 'female secretary', 'road' as 'hard-surfaced road', 'John's book' as 'the book John read/wrote/borrowed'); *conjunction buttressing*, where a conjunction is interpreted as indicating temporal or causal sequence ('John turned the switch and the motor started' implicates that the switch turning preceded, or caused, the starting, or was done with the intention of causing it); and *conditional perfection*, in which a conditional is read as a biconditional ('If' in 'If you mow the lawn, I'll give you $5' is interpreted as 'if and only if') (Levinson 2000, pp.37–8; the examples are Levinson's).

The I-principle allows us to enrich the content of an informationally minimal utterance by drawing on background knowledge. It might seem that the reliance

---

[5] In this example there is also a separate *scalar* implicature to the effect that John does not know that there is life on Mars.

on background knowledge here undermines the status of I-implicatures as generalized and default. I assume Levinson would reply that the knowledge in question is of stereotypes that are both immediately accessible (hence default) and context independent (hence generalized). Thus applying the I-principle does not involve drawing on knowledge of the specific context of utterance or speculating about the speaker's intentions. (This may not be an adequate reply, however; I will return to this issue in section 4 below.)

Levinson summarizes the third principle, the M-Principle as 'What's said in an abnormal way isn't normal' (2000, p.38). This is the reverse of the I-principle: The use of an untypical, or 'marked', expression indicates that the thing referred to is itself atypical in some way. Levinson notes that the principle is related to Grice's maxim of Manner ('Be perspicuous') and in particular to its first and third[6] submaxims: 'avoid obscurity' and 'avoid prolixity'. In flouting these submaxims by using unusual or long-winded expressions, speakers indicate that there is something unusual about the thing described. For example, 'Bill caused the car to stop' (Levinson's example) M-implicates that Bill stopped the car indirectly rather than by simply pressing the footbrake, and 'Jack talked and talked' M-implicates that Jack talked at unusual length. As with Q-implicatures, M-implicatures depend on an implied contrast, this time with unmarked expression that could have been used instead ('stopped the car', 'talked').[7]

Levinson notes that there can be conflicts between the three principles. For example, the Q-principle and the I-principle pull in opposite directions: the former tells us that if a speaker doesn't say something then it should be ruled out; the latter that if a speaker doesn't say something then it can be taken for granted. Similarly, the I-principle tells us to adopt standard interpretations, the M-principle to look for non-standard ones. Levinson argues that these potential conflicts are resolved by assigning priorities to the different types of implicature: Q-implicatures and M-implicatures take priority over I-implicatures; Q-implicatures take priority over M-

---

[6] Levinson actually calls it the fourth, but this seems to be a slip (2000, p.38)

[7] As Levinson notes, the M-principle and the Q-principle both involve *negative* inferences: in both cases the hearer infers the implicated message from the fact that the speaker has *avoided* using some other expression, informatively stronger in one case, less marked in the other (2000, p.40).

implicatures, and clausal Q-implicatures take priority over scalar Q-implicatures.[8] He suggests that all applicable principles are applied automatically, and any inconsistent results subsequently filtered out in accordance with the rules of priority (2000, pp.161–2). Levinson also allows that a Q-implicature can be implicitly cancelled if it conflicts with an entailment of what the speaker says, or is inconsistent with shared background assumptions, or is obviously irrelevant to the speaker's conversational goals (in the last case considerations of relevance, in Grice's sense, will play a role). (Levinson 2000, p.49–52).[9]

1.3 Applying the principles

Levinson gathers a huge amount of data to support the existence of GCIs, highlighting the 'regularity, recurrence, and systematicity' of pragmatic inferences of the kind he describes (2000, p.22). One important piece of evidence comes from facts about lexicalization (Levinson 2000, pp.64–71). English lacks words for the contradictories of certain logical concepts (the concepts are not *lexicalized*). For example, we have a word for *all*, but none for its contradictory *not all*. This, Levinson, argues, is because that concept is carried by 'some' (the contrary of the contradictory of 'all') in virtue of a Q-implicature. The same goes for several other concepts that stand in similar logical relations; for example:

---

[8] Levinson suggests that the priority of Q- and M-implicatures is due to the fact that they involve a deliberate choice of words (a weaker term or a marked expression) rather than reliance on stereotypical interpretation, and that the priority of Q-implicatures over M-implicatures reflects the greater importance of informational content over nuances of expression (2000, p.161).

[9] Other neo-Gricean theorists propose closely related taxonomies; for a useful table comparing them, see Levinson 2000, p.41. In particular, Horn reduces the principles to two: the Q-principle and the R-principle (for example, Horn 2004). The former combines Grice's first submaxim of Quality (be as informative as required) and two submaxims of Manner (avoid obscurity and avoid ambiguity), and it does the combined work of Levinson's Q- and M- principles. The R-principle combines Grice's second submaxim of Quantity (do not be more informative than required), maxim of Relation (be relevant), and third and fourth submaxims of Manner (be brief and be orderly), and it corresponds to Levinson's I-principle.

Not require is unlexicalized since Q-implicated by 'permit'.

May not is unlexicalized since Q-implicated by 'may'.

Not always is unlexicalized since Q-implicated by 'sometimes'.

Not necessary is unlexicalized since Q-implicated by 'possible'.

Not both is unlexicalized since Q-implicated by 'or'.

Levinson notes that these patterns arise in other languages too, suggesting that they are due to the operation of a general interpretative principle (2000, p.69).

Levinson reviews the principles and their application in great detail, showing how they can explain a wide variety of linguistic features and intuitions and raising and responding to numerous objections. He also argues that GCIs are deeply involved in processes of disambiguation, indexical resolution, reference identification, ellipsis unpacking, generality narrowing, and co-reference (anaphora), which are necessary to establish the truth-conditional content of an utterance, and have been traditionally thought of as part of semantic processing.[10] As he notes, this creates a problem for Grice's view that implicatures are determined in part by what is said (the truth-conditional content expressed), since what is said may itself be determined by implicatures. (Levinson calls this 'Grice's circle'; 2000, p.186). Levinson himself avoids this problem by arguing that the GCI principles can be applied to utterance fragments (words or phrases), before a complete propositional content has been determined. This view does, however, present a challenge to the traditional conception of the relation between semantics and pragmatics, on which semantic processing yields a fully-fledged propositional content, which is then enriched or supplemented by pragmatic processes. The upshot, Levinson suggests, is that there are two rounds of pragmatic processing —

---

[10] Levinson also highlights the importance of *intrusive constructions*, such as comparatives and conditionals, where the truth conditions of a sentence depend on an implicature generated by a part of it (Levinson 2000, pp.198–217). An example is 'Driving home and drinking three beers is better than drinking three beers and driving home', where the proposition expressed by the whole sentence is determined by I-implicatures of temporal sequence generated by its two component phrases.

a presemantic round, which establishes a truth-conditional content and a postsemantic round, which may produce a further implicature (Levinson 2000, pp.187–8).

2. Alternatives to neo-Griceanism

In assessing neo-Griceanism, it will be helpful to compare it with rival theories of implicature recovery, and I will briefly introduce three of these in this section. As we shall see, there are reasons for thinking that each has some advantages over neo-Griceanism, and it may be that a theory of implicature recovery can draw on elements from all of them.

2.1 Relevance theory

In the linguistic literature, the neo-Gricean approach to implicature recovery is usually contrasted with that of *relevance theory* (for example, Carston 2002; Sperber and Wilson 1995; Wilson and Sperber 2004). The theory (really a cluster of closely related theories) is complex and has developed over time.[11] Here I shall give a simplified outline, emphasizing the contrast with the neo-Griceanism.[12]

According to relevance theorists, a hearer infers a speaker's meaning from the linguistically coded meaning of their words and contextual information, searching for the interpretation that is the most *relevant* one, in a certain technical sense. The relevance of an utterance is a measure of its *positive cognitive effects* (in particular its *contextual implications* — conclusions one can draw from it in the context), set against the *effort* it takes to process it. Relevance theorists hold that human cognition is automatically geared to maximize the relevance of the inputs it receives, aiming for maximum effects for minimum effort.

---

[11] As Wayne Davis puts it,

> Exposition [of relevance theory] is difficult because formulation of the theory varies significantly from presentation to presentation. And many interlocking technical terms require considerable clarification.' (Davis 1998, p. 99)

[12] The following outline draws in particular on Wilson and Sperber 2004 and Carston 2004a, 2004b.

Since speakers want hearers to attend to what they say, utterances carry a presumption of *optimal relevance* — that is, that they are both (a) sufficiently relevant to be worth the hearer's effort to process them and (b) the most relevant the speaker is able and willing to provide. This presumption (the 'communicative principle of relevance') gives hearers specific expectations of relevance and guides how they interpret utterances. A hearer seeks to infer the speaker's meaning from their words and the context of the utterance, forming and testing hypotheses until their expectations of relevance are satisfied. Note that since clause (b) refers to the speaker's abilities and preferences, relevance theory, unlike the Gricean framework, treats interpretations as sensitive to information about the particular speaker, and it can allow for the possibility that speakers are uncooperative (see, for example, Carston 1998).

According to relevance theory, the interpretation process starts with the linguistically coded content of the utterance (roughly, the context-independent meaning of the sentence uttered), which will typically be underspecified and fail to express a propositional content. Interpreting the utterance then involves two tasks. First, there is a process of what relevance theorists call *explicature*, which involves enriching the linguistically coded content to produce an explicit propositional content, corresponding to what the speaker literally meant. Relevance theorists argue that this process involves not only resolving ambiguities and identifying references, but also a substantial process of pragmatic enrichment, including filling in missing conceptual elements (for example, expanding 'It's raining' to 'It's raining in Sheffield') and narrowing down (or broadening) the meaning of expressions to express more specific ad hoc concepts (for example, narrowing down the meaning of 'happy' to express some contextually salient level or type of happiness).[13] Second, there is a search for additional implicated

[13] From a relevance theory perspective, talk of the 'literal' meaning of an utterance is thus ambiguous. It might refer either to its linguistically coded content as opposed to an enriched explicature of that content, or it might refer to the explicature of the utterance as opposed to a distinct implicature of it. Moreover, neither the linguistically coded content of an utterance nor the explicature of it correspond to what is said by it, in Grice's sense. What is said is richer than the linguistically coded content. (Grice allows that determining what is said requires resolving

meanings (implicatures) distinct from the explicit meaning. Crucially, relevance theorists hold that these processes follow a path of least effort, starting with the simplest, most accessible interpretation and progressing to more complex ones only if current expectations of relevance have not been met. (Although the most accessible interpretation of an expression will usually be what we would regard as the literal one, it may not always be so. Sometimes the linguistic context may strongly prime for a pragmatically enriched meaning, and sometimes an enriched meaning may be much easier to process, as with some metaphors; Noveck and Sperber 2012, p.371.)[14]

---

ambiguities and identifying references; Grice 1975/1989, p.25). Yet what is said is weaker than the explicature, since it does not depend on pragmatic enrichment. We might say that what is said by an utterance is the *minimal proposition* it expresses — the minimal filling in of its linguistically coded content needed to generate a propositional content (Recanati 1993). One problem for this view, however, is that such a minimal proposition will often be quite different from what the speaker means, and will often be trivially true or trivially false. (Consider, for example, 'Everyone screamed' and 'It's snowing', which, without further specification of the relevant domain, will always be respectively false and true.) Since Grice holds that what is said must be meant by the speaker (1968/1989, p.88), this is an implausible consequence. For more discussion of this tricky topic, see Carston 2004b (from where the examples just given are taken) and for a useful table comparing different theorists' use of 'what is said', 'explicature', 'implicature', and related terms, see Levinson 2000, p.195. Saul defends the Gricean notion of what is said, arguing that it is normative rather than psychological (Saul 2002b). As noted in the previous section, Levinson also holds that pragmatic processes contribute to fixing the truth-conditional content of utterances. However, he holds that these processes are limited to application of the GCI principles (as opposed to context-specific enrichment), and he does not recognize a distinction between explicature and implicature, which, he argues, has no principled basis (2000, pp.194–8).

[14] The searches for explicatures and implicatures should not be thought of as independent of each other. The search for an explicature may be constrained by the need to find an explicit meaning that supports a contextually relevant implicature. For example, if a person replies 'I'm tired' when asked 'Do you want to go out?', then interpreting their utterance may involve narrowing down the meaning of 'tired' to a more specific degree of tiredness suitable to implicate 'I don't want to go

Since relevance theory holds that implicature derivation is guided by psychosocial principles, it is in a very broad sense Gricean (in effect, it puts all the weight of derivation on the maxim of Relevance, reinterpreted as the presumption of optimal relevance in the technical sense). However, there are big differences between it and Levinson's neo-Gricean theory. First, Levinson holds that the GCI principles are applied at an early stage in language processing and that the interpretations they yield are the default ones. Relevance theory, by contrast, holds that meanings are processed in order of accessibility, starting with their semantically coded content and enriching it (and deriving implicatures if necessary) until current expectations of relevance are met. As we shall see later, this difference between neo-Griceanism and relevance theory yields conflicting predictions, which have been experimentally tested. A second difference is that neo-Griceanism holds that some implicatures are generalized and context-independent, whereas relevance theory sees implicature derivation as *context-driven* (for example, Breheny et al. 2006). Since optimal relevance is defined in terms of the speaker's abilities and preferences, hearers' expectations of relevance will vary from context to context, and similar utterances may generate an implicature in one context but not in another.

To illustrate this, consider the following examples (taken from Sperber and Wilson 1995, p.277):

(2) *Henry*: If you or some of your neighbours have pets, you shouldn't use this pesticide in your garden.
Mary: Thanks. We don't have pets, but some of our neighbours certainly do.

(3) *Henry*: Do all, or at least some, of your neighbours have pets?
Mary: Some of them do.

---

out'. Interpretation involves a search for the *combination* of explicit and implicit contents that together makes the utterance optimally relevant (Noveck and Sperber 2012, p.313).

Here, neo-Griceanism predicts that both of Mary's replies should implicate that not all of her neighbours have pets, thanks to an automatic application of the Q-principle. Sperber and Wilson suggest that this is wrong; only her second reply carries that implicature. In the first example, the reading of 'some' as 'some and possibly all' (which Sperber and Wilson assume is the more basic one) is sufficient to satisfy Henry's expectations of relevance. In the context it does not matter whether all of Mary's neighbours have pets. In the second example, by contrast, Henry has made it clear that it is relevant to him to know whether all of Mary's neighbours have pets, and Mary's answer would not meet this expectation on the basic reading of 'some'. Henry therefore engages in further processing, reasoning that Mary did not say that all of her neighbours had pets because she was not in a position to do so, and that she means him to understand that they do *not* all have pets (Sperber and Wilson 1995, pp.277–8).

To sum up, relevance theory holds that there are no general inferential principles involved in the derivation of implicatures (other than the presumption of optimal relevance), and no distinction between generalized and particularized implicatures; in effect, it treats all implicatures as particularized.

2.2 Convention theory

In the literature on implicature recovery, neo-Griceanism and relevance theory are the main players. However, I want to introduce another approach, which draws on a non-Gricean analysis of conversational implicature developed by Wayne Davis (Davis 1998; see also Morgan 1978).

Davis argues that conversational implicatures cannot be calculated, even in principle, by applying Gricean conversational principles. I have discussed some of his arguments in previous chapters (for example, Chapter 3, sections 1.1–1.3, Chapter 4, section 1.3), and I shall introduce another one later in section 3.2 below. In opposition to Grice, Davis argues that particularized implicatures (he calls them 'speaker implicatures') are determined by the intentions of the speaker, and that generalized implicatures ('sentence implicatures') are determined by semantic conventions. It is this latter claim that I want to focus on here.

According to Davis, languages are associated with implicature conventions. It is a convention of English that sentences of the form 'Some F are G' are used to

implicate that not all F are G, that sentences of the form 'p or q' implicate 'Not both p and q', that sentences of the form 'p and q' implicate that p preceded q, and so on.[15] Davis defines a convention as

> *an arbitrary social custom or practice*. More explicitly, a convention is *a regularity in the voluntary action of a group that is socially useful, self-perpetuating, and arbitrary*. (1998, p.133; italics in original)

And he argues that implicature practices of the sort just mentioned are conventions in this sense. They are socially useful, promoting '*cooperative, efficient, polite, and stylish communication*' (1998, p.174, italics in original). It is often quicker, politer, and more stylish to implicate something than to say it explicitly. They are largely arbitrary; different implicature practices are possible, and it is a historical accident that we have the ones we do.[16] (Davis accepts that there will have been some pre-existing relation between the literal meaning of a sentence and the implicature it has come to carry, which made the practice 'fitting', 'appropriate', or 'intelligible' (he calls this the *Principle of Antecedent Relation*) (1998, pp.183–4). However, he argues this relation is never strong enough to uniquely determine the implicature.) Finally, implicature practices are self-perpetuating; once a

---

[15] Davis also argues that there are more general implicature conventions, which are not associated with a particular sentence form, but are in effect procedures for generating one-off speaker implicatures. For example, we have conventions of implicating one thing by asserting its denial (as in irony); of implicating a piece of information by making a closely related statement (as in Grice's example of saying 'There is a garage round the corner' to convey where petrol can be bought; or of implicating an affirmative or negative answer by asking a question whose answer is obvious (as in 'Is the pope Catholic?') (Davis 1998, pp.148–54; see also Morgan 1978).

[16] Davis also points out that some generalized implicatures are language-specific, citing Wierzbicka's work on cross-cultural pragmatics (1985, 1987, 1991). For example, in English 'An X is an X' implicates that one X is as good as another, but the Polish equivalent implicates that there is something uniquely good about an X (Davis 1998, p.144).

practice has become established, people have reason to continue following it if they wish to communicate successfully.[17]

Davis claims that the semantic conventions that fix implicatures are of a different type from those that that fix literal meaning (sentence meaning, or conventional meaning in the usual sense). The latter are *first-order*: they are rules for assigning meanings to words and sentences. By contrast, the conventions that determine implicatures are *second-order*: they are rules for assigning further meanings to sentences when they are used with the meanings assigned by the first-order conventions. Thus, first-order conventions dictate that 'Some politicians take bribes' means that some politicians take bribes, and a second-order convention dictates that using that sentence with that meaning expresses the *further* meaning that not all politicians take bribes. As Davis puts it:

> The first-order rules are conventions for using sentences to *directly* express certain thoughts. The second-order rules are conventions for *indirect* expression, rules for expressing further thoughts by expressing thoughts assigned by first-order rules. (Davis 1998, p.156)[18]

Davis holds that languages are defined by their first-order rules, not their second-order ones and that a language's implicature conventions are not essential

---

[17] Morgan also argues for the existence of implicature conventions, though from a Gricean perspective (Morgan 1978). Take the use of 'Can you X?' to request that the hearer do X. Originally, Morgan suggests, this implicature was a particularized one, and the hearer inferred the connection between the literal and implicated content by applying Gricean principles. But as the use of such indirect requests spread, a convention was established whereby one could request someone to do X by saying 'Can you X?', and hearers no longer needed to calculate or even notice the rational connection. The implicature became, as Morgan puts it, 'short-circuited'. Although Davis denies that implicatures are calculable in the Gricean way, he envisages a similar historical process, in which repeated use of a one-off implicature gradually establishes a conventional connection between literal and implicated content (Davis 1998, pp.164–5).

[18] This distinction corresponds closely to Searle's distinction between *conventions of language* and *conventions of usage* (Searle 1975; see also Morgan 1978).

to it. He compares implicature conventions to *speech act rituals*, such as saying 'N speaking' when answering the telephone or asking 'How are you?' when greeting someone. Because implicature conventions are second-order, Davis predicts that second-language learners should take longer to master them than lexical conventions, at least when they are different from those of their first language (Davis 1998, pp.161–2).

Davis also contrasts sentence implicatures with idioms, such as 'kicked the bucket' (meaning 'died') (1998, pp.162–6). Like sentence implicatures, idiomatic meanings are not derivable from the literal meaning of the component words, though there is some relation between the literal and idiomatic meaning that makes the connection appropriate. However, unlike sentence implicatures, idioms do not depend on the current literal meaning of the words used. 'Kicked the bucket' does not mean 'died' in virtue of meaning 'struck the bucket with the foot'. Davis suggests that idioms typically start life as nonce implicatures (metaphors), which later become conventional and finally fossilized into idioms. The literal meanings dropped out of the picture, and the phrases came to express the idiomatic meanings directly. (In effect, second-order conventions became first-order ones.)[19]

I shall refer to Davis's view of generalized implicature as *convention theory*. As will be clear from the previous summary, convention theory is a theory of implicature generation — of what makes it the case that certain utterances carry certain implicatures. There is therefore no *direct* comparison between it and either neo-Griceanism or relevance theory, which are (primarily) theories of implicature recovery. However, as we have seen, issues of generation and recovery are closely interconnected and convention theory does have some implications for implicature recovery. (I should stress that these implications are not identified by Davis himself, who focuses on issues of implicature generation. My remarks here and

---

[19] As noted in Chapter 2, Grice also holds that some implicatures are conventional. For example, 'p therefore q' conventionally implicates that q follows from p (Grice 1989, pp.25–6). However, these implicatures are different from sentence implicatures in Davis's sense. They are part of the meaning of the words used and depend on first-order conventions. Davis's claim is that many of what Grice regarded as *non-conventional*, 'conversational' implicatures are also conventions, though of a second-order kind (Davis 1998, p.157).

later in this chapter may be thought of as a preliminary sketch for a cognitive counterpart to convention theory.)

First, if convention theory is correct, then recovering a generalized implicature will involve knowing and applying the relevant second-order convention, as opposed to applying a general inferential principle or searching for an optimally relevant interpretation. Exactly what cognitive states and processes are involved in this is of course an open empirical question. Second, like neo-Gricean theory, convention theory holds that some implicature processing at least is not context-driven, but involves applying context-independent rules. Third, convention theory agrees with relevance theory that literal (linguistically coded) meanings are in a sense psychologically more basic than implicated ones. As we have seen, knowledge of literal meanings requires mastery of first-order conventions only, whereas knowledge of implicatures requires mastery of second-order conventions as well. Since it is possible to acquire the former without the latter (as in the case of second language learners), this suggests that knowledge of lexical conventions may be stored and accessed separately from knowledge implicature conventions. As we shall see in section 5, this means that convention theory may offer new ways of interpreting experimental work on scalar implicature.

2.3 Weak neo-Griceanism

The fourth approach to implicature recovery that I want to introduce is one that, so far as I know, does not have a name, though it is a possible position and one that is represented in the literature. (I shall consider an example later.) It is an inclusive position, which can be seen as a halfway house between neo-Griceanism and relevance theory.

Neo-Griceanism can be thought of as involving two core claims: (1) some implicatures are derived from the application of broadly Gricean principles, and (2) these principles are applied *by default*. Claim (2) itself can be understood to mean (a) that the principles are applied automatically whenever an appropriate expression is detected, regardless of context, and (b) that they yield the same interpretations of the same expressions each time. (Subclaim (b) follows from subclaim (a); it is because the principles are not sensitive to context that they yield the same interpretations each time.) Claim (2) is important to Levinson's view that

the principles serve to speed up language processing; it is because they provide rapid context-independent enrichments that they save time. But (1) does not entail (2), and the principles might still serve a useful interpretative function even if they are not applied by default.

The view I want to introduce accepts (1) but rejects (or remains neutral about) (2). That is, it holds that we derive some implicatures by applying Gricean principles, such as the Q-principle, but does not claim that we do so by default. It allows that contextual factors may determine when and how the principles are applied, and that the principles may yield different interpretations of the same expressions in different contexts. This view agrees with neo-Griceanism that some implicatures are derived by applying general pragmatic principles but holds that the principles in question are applied in a context-sensitive way — perhaps when a literal interpretation fails to meet relevance expectations. Hence, it agrees with relevance theory that there are no truly generalized, context-independent implicatures. I shall refer to this broad approach as *weak neo-Griceanism* ('Griceanism' because its principles are derived from Grice's; 'neo' because it applies Gricean principles to implicature recovery; and 'weak' because it is not committed to (2)).

2.4 Back to Levinson

Having outlined these broad alternatives to neo-Griceanism, I shall now turn to the task of assessing neo-Griceanism itself, in the form developed by Levinson. I shall highlight some problems for the theory and indicate how one or other of the alternative approaches might be applied instead. I shall not attempt to adjudicate between the alternatives themselves but merely show that each may better explain some of the cases discussed.

As already noted, Levinson discusses a huge number of examples, often persuasively, and I cannot possibly engage with the range and detail of his analyses. (Nor, indeed, can I consider more than a tiny fraction of the alternative analyses offered by the rival approaches.) But I will focus on some key problem cases.

3. Assessing the Q-principle

This section looks at some problems relating to Q-implicatures. It looks first at a core example, and then discusses some more general theoretical concerns about the neo-Gricean treatment of scalar implicatures.[20]

3.1 'An X'

Since Neo-Griceanism is derived from Grice's account of generalized implicature, I will begin by returning to the example Grice uses to introduce the notion (which he describes, in a characteristically cautious way, as one that he 'hope[s] may be fairly noncontroversial'; Grice 1989, p.37). This example, which I discussed briefly in Chapter 2, concerns the indefinite article. Grice notes that when a speaker uses an expression of the form *an X*, they typically implicate that the X in question is not closely connected to them. For example, 'I met a man' implicates that the man was not my close relation or friend, 'I found a cat' implicates that the cat was not mine or one known to me, and so on. Grice proposes that this is due to the first maxim of Quantity:

> the implicature is present because the speaker has failed to be specific in a way in which he might have been expected to be specific, with the consequence that it is likely to be assumed that he is not in position to be specific. (Grice 1975/1989, p.38)

Grice explains that when a person or object is familiar to the speaker, it will usually be informative to indicate that it is, since our interactions with familiar persons and objects are typically very different from our interactions with unfamiliar ones. If a

---

[20] One common area of application for the Q-principle is in relation to number words. Neo-Griceans typically hold that these literally specify only minimum amounts ('three' literally means 'at least three') and that the exact meanings they commonly carry ('exactly three') are the product of implicatures generated by application of the Q-principle. However, number words present special problems (see, for example, Carston 1998; Levinson 2000, p.88), and I shall not focus on them here.

speaker does not indicate that the person or item in question was familiar (for example, by using ‘my’ rather than ‘an’), they implicate that it was not familiar.

There is a problem with this, however. For, as Grice notes, in some cases the implicature does not hold. ‘I have been sitting in a car all morning’ does not implicate that the car was not my own. And in some cases the opposite implicature holds. ‘I broke a finger yesterday’ implicates that the finger was mine. (Both examples are Grice’s.) Since there is no explicit cancellation in these cases, how can ‘an X’ carry a generalized implicature of lack of connection?

Levinson revisits this case and proposes a slightly different and more comprehensive account. He points out that the implicature from ‘an X’ to ‘not my X’ cannot be a simple Q-implicature, since possession and indefiniteness are different types of relation, and there is therefore no scale <*my, a*>. Rather, he argues, it is a two-stage process, involving a Q-implicature followed by an I-implicature. First, there is an entailment scale <*the, a*> since definite and indefinite reference are similar relations, so ‘an X’ Q-implicates ‘not the X’. In using the indefinite article, we implicate that we do not mean to refer to some definite, unique X. Second, ‘the X’ I-implicates ‘the salient X’ — that is, the one I am familiar with or closely connected to in some way.

> Grice’s examples are thus indirectly explained: when one says ‘I went into a house’ one Q-implicates ‘I didn’t go into the house’, where the definite suggests (I-implicates) my very own house. (Levinson 2000, p.92)

This is neat. But how is ‘I broke a finger’ to be explained? According to this account, it should implicate that the speaker cut someone else’s finger, not their own, but that would not be the usual interpretation. Levinson agrees but argues that this is not a counterexample. He explains (taking ‘I cut a finger’ as his example):

> [W]hen I say ‘I cut a finger’ I merely implicate that it was no unique, otherwise salient finger (say, the one I cut before) that suffered, and that interpretation is compatible with the assumption that it was my own

> finger (which in turn is a more stereotypical reading than one that involves the chopping of other peoples' fingers). (Levinson 2000, p.92)

The idea is that the Q-implicature to 'not *the* finger' goes through as usual, indicating that the speaker did not mean to pick out some specially salient finger, such as the one that they had been planning to cut or had cut the day before. But this does not rule out the finger being his or her own.

There are problems with this response, however. First and most obviously, although the implicated message is compatible with the claim that the finger in question was one of the speaker's own, it is also compatible with the claim that the finger was someone else's. If a manicurist were talking, for example, it would be natural to read the utterance as implicating that they had cut a client's finger — as Levinson himself acknowledges (2000, p.17). Yet the implicated message is not that the finger *might* have been my own, but that it *was* my own. The parenthesis in the passage just quoted suggests that Levinson would respond by appealing to I-principle. If the speaker is not a manicurist, then the stereotypical reading will be one on which the finger was their own. But this cannot be right. For the I-principle (we are assuming) tells us to draw on *general* background knowledge, not on knowledge of the specific context. This would include knowledge that manicurists use sharp objects on other people's fingers, whereas non-manicurists rarely do this, but it would not include knowledge that the speaker himself is a manicurist. That is context-specific knowledge and hence irrelevant to GCIs. So the implicature that the finger cut was someone's else cannot be a *generalized* one. At most there is a generalized implicature that it was not the salient one — whichever that might be.

A second problem for Levinson's account is that it does not explain other very similar cases. In English, 'I cut a finger' implicates that the cut finger was the speaker's own, but the formally and semantically similar 'I cut a nose' typically carries the opposite implicature — that the nose was *not* the speaker's own. On Levinson's account it is hard to see why this should be so. Use of 'a nose' should implicate 'not *the* nose' — the unique, otherwise salient nose — and that is compatible with the assumption that it was the speaker's own. Moreover, an appeal to stereotypicality reinforces this. We are more likely to cut our own noses than to cut other people's, just as we are more likely to chop our own fingers than those

of others. It seems that other factors or principles must be in play here, beyond those mentioned by Levinson. In fact, the uniqueness or otherwise of the bodily part in question seems to be crucial here. 'I hurt an X' implicates that the X is the speaker's own if the speaker has more than one X (for example, finger, toe, ear, breast, testicle), but implicates 'The X was someone else's' if the speaker has just one X (for example, head, nose, chin, penis, vagina, etc.). We might call this the *Uniqueness Principle*. Why should this hold? It might be suggested that since 'an X' implicates 'not the X', it cannot refer to a unique feature of the speaker. Compare 'A wheel of the car is loose', where the use of 'a' rather than 'the' indicates that the reference is to a (non-unique) road wheel rather than the (unique) steering wheel (example adapted from Levinson 2000, p.155). The cases are not parallel, however. A speaker would not refer to their own nose as 'the nose', and 'nose' unlike 'wheel' is not ambiguous between unique and non-unique features. Besides, this does not get to the heart of the matter. The implications we need to explain are not of uniqueness or otherwise, but of *possession*. Why should 'a nose' implicate 'not my nose' while 'a finger' implicates 'my finger'? It is not obvious that this can be explained with the resources Levinson has to offer, and a more plausible explanation may be that it is a simply a convention of English usage, in line with convention theory.[21]

It turns out then, that Levinson's account of the 'an X' case, does not fare much better than Grice's. If GCIs are genuinely context-independent, then the same expression-form should give rise to the same implicature in every context (and indeed every language), but 'an X' does not, and neither Grice nor Levinson can fully explain why. Given that this was the example with the notion of generalized implicature was originally introduced into the literature, this is a problem for

---

[21] Informal discussions with speakers of other languages suggest that the Uniqueness Principle is indeed a convention of English. In many languages it is not applicable, since it not felicitous to use a 'a leg/nose' in this context without explicitly indicating whose it is, by use of a pronoun or reflexive verb. However, in languages where the phrase is not infelicitous, the principle does not always hold. In Finnish, for example, the reference would be to the speaker's own X, whether unique or not, and in Greek the reference would be ambiguous. (My thanks to the friends and correspondents who have shared their intuitions on this topic with me.)

Gricean approaches. Even if there are generalized implicatures associated with use of the indefinite article, they do not appear to be derivable from general principles and may depend on language-specific conventions.

3.2 Scalar implicatures

I turn now to some more general considerations about Q-implicatures, and, specifically, scalar implicatures. These are perhaps the clearest examples of GCIs, and Levinson's account of them builds on a rich body of pre-existing work on the topic, including Gazdar 1979, Horn 1972, 1984, 1989, and Hirschberg 1985. However, even here there are reasons for thinking that the neo-Gricean account may not be the best. I shall begin with a general worry about the Gricean approach to scalar implicature, raised by Wayne Davis (2014).

As Davis points out, the idea behind Quantity implicatures is that we recover the implicated message by reference to what is *not* said. He quotes Levinson's own (1983) account of the implicit reasoning involved:

> (i) S has said *p*
>
> (ii) There is an expression *q*, more informative than *p* (and thus *q* entails *p*), which might be desirable as a contribution to the current purposes of the exchange (and here there is perhaps an implicit reference to the maxim of Relevance)
>
> (iii) *q* is of roughly equal brevity to *p*; so S did not say *p* rather than *q* simply in order to be brief (i.e. to conform to the maxim of Manner)
>
> (iv) Since if S knew that *q* holds but nevertheless uttered *p* he would be in breach of the injunction to make his contribution as informative as is required, S must mean me, the addressee, to infer that S knows that *q* is not the case (K~*q*), or at least that he does not know that *q* is the case (~K*q*).
>
> (Levinson 1983, p.135)

But of course, in any exchange there are *many* things that are not being said. As an example, Davis takes 'Some athletes smoke'. This implicates the denial of the stronger claim that all athletes smoke. But that is far from being the only stronger

relevant statement of roughly equal brevity the speaker could have made and did not. Davis illustrates this with a range of scales on all of which ‘Some athletes smoke’ figures as the weakest element:

> <All athletes smoke, Nearly all athletes smoke, Most athletes smoke, Many athletes smoke, Several athletes smoke, Some athletes smoke>
> <100% of athletes smoke, At least 90% of athletes smoke, At least 50% of athletes smoke, At least 10% of athletes smoke, At least 1% of athletes smoke, Some athletes smoke>
> <Some athletes smoke constantly, Some athletes smoke regularly, Some athletes smoke often, Some athletes smoke occasionally, Some athletes smoke>
> <Some athletes, maids, and cops smoke, Some athletes and maids smoke, Some athletes smoke>
> <Some athletes smoke filterless Marlboros, Some athletes smoke Marlboros, Some athletes smoke>
> <Everyone knows some athletes smoke, I know some athletes smoke, Some athletes smoke>
> <n% of athletes smoke ($0 < n < 100$), Only some athletes smoke, Some athletes smoke>
> (Adapted from Davis 2014)

By Levinson’s reasoning, Davis argues, ‘Some athletes smoke’ should implicate the denial of all of the stronger statements. Yet in fact it implicates the denial of only one of them:

> Among the infinity of statements stronger than ‘Some athletes smoke,’ ‘All athletes smoke' is highly unusual in that people typically implicate its denial. (Davis 2014)

Now Levinson has an answer to this. In the passage quoted above he mentions that the stronger statement must be ‘of roughly equal brevity’ and ‘desirable as a contribution to the current purposes of the exchange’. And in later work he sets

out two more precise constraints that an entailment scale must meet in order to support Q-implicatures. First, the stronger items in the scale must be lexicalized to at least the same degree as the weaker ones. That is, the stronger items must consist of as few or fewer words than the weaker ones, so that, for example, if the weakest item is monolexemic, then all the other elements are monolexemic too. Second, all the items in the scale must be 'about' the same semantic relations and thus 'in conceptually salient opposition'. So for example, the scale <*regret*, *know*> does not support Q-implicatures, since 'regret' involves a conceptual element not present in 'know' (Levinson 2000, p.80). And these conditions rule out most of Davis's examples. None of his scales except the first meets the lexicalization constraint, and the penultimate one at least fails the aboutness constraint.

The lexicalization constraint also gives Levinson a response to what would otherwise be a serious objection. He holds that 'and' is typically strengthened by application of the I-principle to indicate temporal or causal sequence (2000, p.37–8). So 'They got married and had a child' I-implicates that the marriage preceded the child's birth. But it seems that the Q-principle could also be applied here to produce precisely the opposite implicature. Since the speaker chose 'and' rather than the informationally stronger 'and then', we might infer that they were not in a position to assert that the events took place in that order described and thus that they do not know that they did and perhaps know that they did not. Since Q-implicatures take precedence over I-implicatures, we should therefore take the utterance to implicate that the child's birth did not take place after the marriage (Davis 1998, p.52–3). However, the scale involved here, <*and then*, *and*>, does not satisfy the lexicalization constraint, since the stronger element is less lexicalized than the weaker one, and it does not therefore support Q-implicatures (Bezuidenhout 2002, p.264–5).[22]

---

[22] We might press the objection, pointing out that some speakers use 'then' (incorrectly, according to grammarians) as a coordinating conjunction, as in 'They got married, then had a child', which suggests that in their idiolect at least there is a legitimate scale <*then*, *and*> which does support the Q-implicature. However, it may be that in these cases the comma before 'then' serves as a coordinating conjunction, so the real scale is <[comma] *then*, *and*>, which, arguably, does not meet the lexicalization condition.

But why should the Q-principle be restricted in this way? (The fact that it saves Levinson's account from a serious objection is not itself a reason, unless one is already convinced that the account is correct.) If we typically assume that what is not said isn't the case, then why doesn't saying that some athletes smoke implicate that it is not the case that some athletes, maids, and cops smoke? What justifies the two constraints? Levinson suggests answers. He justifies the lexicalization constraint on the grounds that where stronger items are *less* lexicalized (more wordy), then any Q-implicature would be undercut, since the hearer might think that the speaker had avoided the stronger term simply because they were avoiding being 'clumsy and prolix' (following Grice's maxim of Manner), rather than because they were not in a position to assert it (Levinson 2000, pp.79–80). I assume the aboutness constraint is justified in a similar way, by reference to the maxim of Relevance. Where a stronger term would have introduced a different kind of information, the hearer may take the speaker to have avoided it simply in order to remain relevant, rather than because they were not in a position to assert it, thus undercutting any potential Q-implicature it might have supported. These justifications are not unreasonable, and plainly the Q-principle (if it is a principle) would have to be restricted in some way — otherwise it would produce endless implicatures from every utterance. However, we might wonder if an appeal to Gricean maxims is sufficient to justify the conditions in the strict form Levinson proposes. The assumption that a speaker is following the maxims of Manner and Relevance doesn't require us to suppose that they would have avoided even *slightly* longer phrases or introduced *any* new information. Moreover, even if we accept the conditions, some problems remain.

First, the constraints do not rule out all the problematic scales. For example, consider the scale <*several*, *some*>. This meets the two constraints, but does not support GCIs. 'Some athletes smoke' does not imply that it is not the case that several athletes smoke. Levinson might reply that 'several' forms part of a larger scale that continues up to 'all' and that it is only the strongest element on the scale whose denial is implicated (Levinson 2000, p.77). However, even if this further qualification is added, it is arguable that exceptions remain. For example, consider:

<*is a cardiologist*, *is a physician*>

<commit murder, commit a crime>

<likes baseball, likes sports>

Again, these scales meet Levinson's constraints but do not support generalized scalar implicatures. 'X is physician' does not generally implicate that X is not a cardiologist; 'Some cops commit crimes' does not generally implicate that no cops commit murder, 'Y likes sports' does not generally implicate that Y does not like baseball (though there might be specific contexts in which those implicatures would hold).

Second, there are entailment scales which do not meet the constraints but *do* plausibly support GCIs. Consider, for example:

<got a distinction, passed>

<right up to, near>

<got a good look at, saw>

These scales do not meet the lexicalization constraint. Yet my intuition is that they support scalar implicatures. 'Amy passed' implicates that Amy did not get a distinction; 'Bob went near the edge' implicates that Bob didn't go right up to the edge; 'Cal saw the robber' implicates that Cal didn't get a good look at the robber.[23]

Perhaps further constraints could be added to deal with these exceptions, but they would begin to look ad hoc. Moreover, if the Q-principle were supplemented with even more constraints, it would no longer look like a rule that could be applied automatically at an early stage of processing. Applying it would be a complicated business, which would involve checking that multiple conditions hold, and it might slow down communication rather than improve its efficiency. If there are such tight constraints on the application of the Q-principle — with the result that the

---

[23] This is confirmed by the recognized diagnostics for scalar implicatures, summarized by Levinson (2000, p.81). If a scale, <S, W>, supports scalar implicatures, then the following cancelling and suspending phrases should be permissible: 'W and even S', 'Not only W, S', 'W in fact/indeed S', 'W or possibly/even S' and 'W if not S'. This is the case with the three scales mentioned.

exceptions to it hugely outnumber the cases to which it applies — then why posit the principle in the first place? Wouldn't it be more economical to appeal to conventions of use rather than the general principles? Perhaps it is a convention of English (and other languages) that when we use 'some', with its basic meaning of 'at least one', it is understood to implicate 'not all'. Thus, in order to recover the implicature, hearers would simply need to know the convention and recognize that it applies in this case. This is a simpler hypothesis than supposing that speakers have to access a relevant scale, check that the scale meets multiple constraints, and then apply an inferential principle.

This conclusion is reinforced, I think, by another point made by Davis (Davis 2014). If someone asks 'Do any athletes smoke' the response 'Some do' will carry the implicature that not all athletes smoke, but the answer 'Yes' will not. Yet, Davis points out, in the context 'Yes' is logically equivalent to 'Some do' and is a no less cooperative response. Since the two exchanges are informationally equivalent we should expect them to produce the same implicatures, if general principles are at work. If the Q-heuristic produces a scalar implicature in the first case, then it should do so in the second too. The fact that no implicature arises in the second case strongly suggests that a general principle is not involved in either case. Again, it seems more appropriate to appeal to a convention of use, which is associated with particular expressions. There may be a convention of use that 'some' (in the right context) implicates 'not all', but there is of course no convention that 'yes' implicates 'not all'.

3.3 Reducing scalar GCIs to PCIs

There is another way of looking at scalar implicatures, which stresses their continuity with particularized implicatures. Levinson allows that there may be other types of Q-implicature, in addition to those based on entailment scales and clausal contrasts (see Levinson 2000, pp.98–103, from where the examples below are taken). For example, Q-implicatures can be based on *non-entailment* scales, such as <*succeed, try*> (saying that John tried to reach the peak implicates that John did not succeed, even though succeeding does not entail trying). Q-implicatures can also be based on sets of *alternatives*, where the choice of one alternative implicates that the others do not apply (for example, 'The flag is white'

implicates ‘The flag is not white and red’), and on *levels of specificity*, where the use of a more general term implicates that the speaker cannot be more specific (for example, ‘I just saw a horrid animal in the larder’ implicates that the speaker is not sure what sort of horrid animal it was). Levinson notes that many of the inferences underlying these implicatures are weak unless contextually reinforced, and thus lie at the border between GCIs and PCIs (2000, p.103).

Following Fauconnier (1975), Levinson also notes that scalar implicatures can be generated by contingent scales, which depend on our beliefs about the world rather than on the meanings of the terms involved. He offers this example (Levinson 2000, p.104):

(4) He can drive small trucks
Implicature: He can’t drive big ones.

The Q-implicature here depends on the scale <*driving big trucks, driving small trucks*> which is based on knowledge of truck-driving rules and skills (people licensed to drive big trucks are also allowed to drive small ones, but not vice versa).

Q-implicatures can also depend on contextually given *nonce* scales. Levinson quotes the following example (Levinson 2000, p.105, quoted from Hirschberg 1985, p.50):

(5) A: Did you get Paul Newman’s autograph?
B: I got Joanne Woodward’s.
Implicature: I didn’t get Paul Newman’s.

Here the speaker assumes a scale of autograph prestige <*Newman*, *Woodward*>, and by affirming that they secured the lower value item, they implicate that they did not secure the higher-value one.

Julia Hirschberg proposes a systematic treatment of scalar implicatures which includes such contingent and context-dependent ones (Hirschberg 1985; for

discussion, see Levinson 2000 pp.104–8).[24] According to this, scalar implicatures are supported by *orderings* (Levinson calls them 'Hirschberg scales') constructed from a contextually salient set of *values* (expressions) and an *ordering relationship* of some kind (it can be any relation that is salient). For example, the values 'oak', 'maple', 'tree' and the ordering relation *is-a-kind-of* would give the following ordering:

<{*oak*, *maple*}, *tree*> [25]

Note that this ordering is a *partial* one, in that it does not apply to every pair of items in the set. 'Oak' and 'maple' are both ordered with respect to 'tree', but not with respect to each other. In Hirschberg's account, scalar implicatures require only partial orderings to support them. Similarly, the ordering relationships *has-parts*, *has-attribute*, and *has-prior-stage*, might yield the following orderings (examples quoted in Levinson 2000, pp.106-7):

<{*book*, {*chapter 1*, *chapter 2*, …}>
<*Greek*, {*Greek-speaking*, *Greek relatives*, *Greek residency*, *Greek ancestry*}>
<*marriage*, *engagement*, *going-steady*, *dating*>

The rules for scalar implicatures are then as follows. Affirming a lower expression in an ordering (to the right) implicates either that the speaker doesn't believe that a higher expression applies or that they do not know which, if any, does. Thus 'It's a tree' implicates that the speaker does not know which kind of tree; 'I've read Chapter 1' implicates that the speaker hasn't read the whole book, and 'I've Greek relatives' implicates that the speaker is not Greek. By contrast, denying a higher

---

[24] Hirschberg is a computer scientist, and her aim is to develop a formal framework for the representation and calculation of scalar implicatures that could be implemented computationally, rather than to identify the psychological mechanisms involved in human implicature recovery. The framework is detailed and complex and only its broad outlines are relevant here.

[25] Example from Levinson 2000, p.106.

item implicates that the speaker believes that a lower one applies, or may do so. Thus, 'It's not an oak' implicates that it is a tree, 'I haven't read the whole book' implicates that the speaker has read some of the chapters, 'We're not married' implicates that the speaker may be engaged or dating, and so on.

Hirschberg holds that scalar implicatures can also be generated by *unordered* sets of alternatives, such as {*chapter 1*, *chapter 2* …}. Here the rule is that affirming one expression implicates that the others do not apply or are not known to apply, and denying one expression implicates that one of the others may apply. So, for example, 'I've read Chapter 1' implicates that the speaker has not read Chapter 2, and 'I've not read Chapter 1' implicates that the speaker may have read Chapter 2 (Levinson 2000, p.106).

Crucially, Hirschberg extends this treatment to scalar implicatures based on entailment scales. Given the contextually salient expressions 'all' and 'some' and the relation of entailment, we can form the Hirschberg scale <*all*, *some*>. Applying the rules, 'Some came' implicates that not all came, and 'Not all came' implicates that some came. Since these implicatures too are generated from contextually salient orderings, it follows that there is no sharp distinction between PCIs and GCIs, and that Hirschberg's approach reduces GCIs to PCIs.[26]

Levinson rejects this conclusion, of course. Though he concedes that Hirschberg offers a neat treatment of particularized scalar implicatures, he argues that it does not tend to undermine the distinction between GCIs and PCIs, since we can still draw a clear-cut distinction between context-independent scalar implicatures that are based on contrasts in meaning (GCIs) and context-dependent ones that are based on contrasts salient in particular contexts (PCIs):

> The GCI theorist is simply claiming that speakers carry their lexicons on their backs, as it were, from context to context, and it is mutual knowledge of this fact that elevates the Q-heuristics to a default mode of inference. (Levinson 2000, p.108)

---

[26] Hirschberg writes: 'the traditional distinction between generalized and particularized implicature is a false one, an artefact of the inventiveness of analysts — or lack thereof' (Hirschberg 1985, p.42).

This is questionable, however. Levinson assumes that some expressions (such as 'some', 'sometimes', 'possibly') will evoke the same set of contrasting values and the same ordering relation (namely, entailment) in all contexts, thus supporting context-independent scalar implicatures. That is, the scales that support GCIs will be salient in all contexts. But this is questionable. Consider these exchanges, for example:

(6) A: Is it true that she bought ten pairs of shoes yesterday?
B: She bought some.

(7) A: Do you visit her as often as you used to?
B: We visit her sometimes.

In the context of (6) 'some' does not evoke the entailment scale <*all*, *some*> but the nonce specificity scale <*ten*, *some*>, and by using 'some' B implicates that she doesn't know if the more specific figure of ten is correct. Similarly, in (7) the entailment scale <*always*, *sometimes*> is not salient, but instead the nonce scale <*as often as we used to*, *sometimes*> is. Of course, in many contexts where 'some' is used, the familiar entailment scale <*all*, *some*> *would* be salient, but as the examples just given show, there are contexts in which it would not be. Given this, it is more appropriate to think of scalar implicatures as lying on a continuum, from relatively particularized ones, which depend on orderings that are salient only in a few contexts, to relatively generalized ones, which depend on orderings that are salient in many contexts. But since this difference is a matter of degree and there will be a full range of intermediate cases, this view tends to undermine the sharp neo-Gricean distinction between PCIs and GCIs, and, with it, Levinson's case for existence of a distinct level of utterance-type meaning.[27]

[27] This is not to deny, of course, that there are general principles at work in scalar implicature, on Hirschberg's account. There are the rules that affirming a lower-ranked element in an ordering implicates the denial of stronger one, and that denial of a stronger element implicates affirmation

Levinson also objects that Hirschberg's account will overgenerate implicatures, since it imposes no constraints on scalehood (echoing Davis's complaint against him, discussed above) (Levinson 2000, p.107). But this, I think, mistakes Hirschberg's aims. In fact, it is a virtue of her account that it imposes no such constraints. For, as Hirschberg stresses, given the right ordering, *any* expression can generate a scalar implicature.[28] Consider again the examples Davis gives. There is no general implicature from 'Some athletes smoke' to 'It is not the case that some athletes smoke Marlboros', but there are contexts where it arises:

(8) A: Do some athletes smoke Marlboros?
B: Some athletes smoke.

The implicature depends on the contextual salience of the expression 'smoke' and 'smoke Marlboros' and the ordering relation *is-a-specific-form-of*. Affirming that some athletes engage in the general activity of smoking implicates that they do not

---

of a weaker one. But these rules by themselves do not generate implicatures, even when combined with the lexicon. A contextually salient ordering must also be given.

[28] Hirschberg illustrates this with the following example. Each of B's replies generates a different scalar implicature, based on a different implicit ordering or set of alternatives:

A: Did the girl in the red dress spill a diet coke?

a. B: She spilled a diet *pepsi*.

b. B: She spilled a *regular* coke.

c. B: She spilled a *glass of tomato juice*.

d. B: *Jane* spilled a diet coke.

e. B: The girl in the red *slacks* spilled a diet coke.

f. B: The girl in the *green* dress spilled a diet coke.

g. B: The girl in the *green slacks* spilled a diet coke,

h. B: The *boy* in the red dress spilled a diet coke.

i. B: The girl in the red dress *will spill* a diet coke.

j. B: The girl in the red dress *drank* a diet coke.

k. B: The girl in the red dress spilled *the* diet coke.

engage in the more specific Marlboro-involving form of smoking (or are not known to do so).[29]

This suggests another way of looking at the constraints on scalehood that Levinson proposes. I have already suggested that these are inadequate, and I think we can now see why. If Hirschberg is right, then our intuitions about which scales do and do not support scalar implicatures are based on our judgements about contextual salience. We judge that the entailment scale <*all*, *some*> does support implicatures and that the specificity scale <*ten*, *some*> does not because in most contexts the former would be salient and the latter would not. But, as we have seen, there are contexts in which the reverse is the case. We cannot hope to provide general rules of scalehood since there are no truly general, context-independent scalar implicatures. Any rules would have potentially unlimited context-specific exceptions.

Of course this means that the notion of contextual salience has a lot of work to do in Hirschberg's account, and Levinson suggests that this is a major problem for the account:

> All implicatures are made dependent on the contextually salient ordering relation, so we have no account of implicature generation without an account of how this is arrived at. (Levinson 2000, p.107)

In fact, Hirschberg devotes a whole chapter to the discussion of contextual salience, identifying some of the cues that make an ordering salient to a speaker and hearer (syntactic, intonational, semantic, pragmatic, and communication dynamical) and proposing ways in which assignments of salience might be

---

[29] Other examples from Davis's list would generate implicatures based on unordered sets of alternatives:

> A: Is it true that athletes, maids, and cops smoke?
>
> B: Some athletes smoke.

Here the salient set of alternatives is {*athletes*, *maids*, *cops*}, and the affirmation of one element implicates the denial of the others.

formally represented in a computational model (Hirschberg 1985, Chapter 6). It is true that this does not amount to (and is not intended to be) a complete theory of contextual salience (indeed, providing such a theory would be a major achievement for psycholinguistics). But that does not undermine the case for thinking that we need such a theory, and Levinson has not shown that we don't need one.

According to Hirschberg, then, there are genuine scalar implicatures, whose recovery involves the application of general principles (affirmation of a weaker scalar term implicates denial of a stronger one, and denial of a stronger one implicates affirmation of a weaker one). But these principles do not yield default, context-independent interpretations. The *orderings* to which the principles are applied are contextually determined, and the same expression might evoke a different ordering, and hence a different scalar implicature, in different contexts. Thus, Hirschberg's approach (at least as I have interpreted it) is an example of the class of views I called *weak neo-Griceanism*.[30]

3.4 Q-implicature and T-implicature

There is another set of considerations that pose a challenge for Levinson's account of Q-implicature. Levinson's principles, like Grice's maxims and the Cooperative Principle from which they follow, are rooted in the assumption that the aim of communication is the efficient sharing of information. Thus, the Q-principle tells us to assume that speakers will give as much (relevant) information as they can. But as several writers have noted, conversation often has other aims besides the communication of information. Sometimes it is more important to be polite than to be informative. Geoffrey Leech has formalized this idea, proposing a *Politeness Principle*, 'Minimize (other things being equal) the expression of impolite beliefs', which he breaks down into a series of maxims, of Tact, Generosity, Approbation, Modesty, Agreement, and Sympathy (Leech 1983, p.132). Leech notes that

---

[30] The particularized implicature carried by Mr Bronston's reply 'The company had an account there', discussed in Chapter 1, might be regarded as a scalar implicature, dependent on the ad hoc scale <*me*, *my company*>, where the ordering relation is something like *seriousness of holding a bank account in the name of*.

communicative exchanges frequently involve a trade-off between the Cooperative Principle and the Politeness Principle, and he explores in detail the complex pragmatics of politeness. Here I shall focus on a narrow range of cases, in order to highlight a potential problem for neo-Griceanism.

In British English at least it is common to use understatement to convey information or instructions that will be unwelcome to the hearer. We might see this practice as obeying what Leech calls the *Tact maxim*: 'Minimize the expression of beliefs which express or imply cost to other' (Leech 1983, p.132). For example, a manager might tell a subordinate, 'There is a problem with your report', to convey that they are in fact seriously displeased with it. We might call this a *Tact implicature, or T-implicature*. Here are some more examples:

(9) I might not be able to do that.
T-implicature: I won't do that.

(10) I would like to see your passport.
T-implicature: You must show me your passport.

(11) Someone's eaten the icing off the cake.[31]
T-implicature: You have eaten the icing off the cake.

Note that these implicatures are not M-implicatures. The expressions used are not untypical or unusual. Indeed, these sentences could be used without generating the implicatures. If (9) were uttered by a friend who was clearly anxious to help as much as they could, or (10) by someone known to be interested in the design of passports, the implicatures would not arise. To this extent, then, these are context-dependent, particularized implicatures. However, they are not *highly* contextualized, and the examples given can be easily understood without any background information.

In these cases, the hearer applies what we might call a *Tact principle:* interpret the affirmation of a weaker, less unwelcome statement or request as implicating a

[31] Example borrowed from Leech 1983, p.80.

relevant stronger, more unwelcome one. What isn't said, is, we might say. This is, of course, the opposite of Levinson's Q-principle, which says that affirmation of a weaker claim implicates the denial of a relevant stronger one. Moreover, the same utterance may potentially generate both a Q-implicature and a T-implicature. In (9), 'I might not be able' should Q-implicate that it is not the case that the speaker definitely won't be able to do the thing requested — which is, of course, at odds with the T-implicature that the speaker won't do it. Here are some more examples:

(12) It is possible that the train will be delayed.
Q-implicature: It is not probable that the train will be delayed.
T-implicature: It is probable that the train will be delayed.

(13) Some of the staff you sacked are angry.
Q-implicature: Not all the staff you sacked are angry.
T-implicature: Many or even all of the staff you sacked are angry.

(14) I think you dropped this.
Q-implicature: I am not sure that you dropped this.
T-implicature: You dropped this.

Given the right context, the T-implicatures here would take precedence over the Q-implicatures. (Imagine (12) said by a grim-faced railway employee.) And this poses a problem for Levinson's account, which holds that Q-implicatures are generalized and take priority over other implicatures. It is true, as I mentioned earlier, that Levinson allows that Q-implicatures can be cancelled if they conflict with entailments of what is said or with background assumptions, or if they are obviously irrelevant. But it would be a major concession to allow that Q-implicatures can also be overridden by T-implicatures, which are context-dependent, play a social role rather than an informational one, and may even be culturally determined.

This is only tentative, of course, but it tends to support the earlier suggestion that scalar implicature is much more context sensitive than Levinson allows. I

suspect this point could be reinforced by considering other manifestations of the Politeness Principle.

3.5 Scalar implicature or explicature?

I shall close this section by introducing another alternative to the neo-Gricean treatment of scalar implicature, this time from a relevance theory perspective. (The following draws on Noveck and Sperber 2012.)[32]

The view in question is that some supposed scalar implicatures are not in fact implicatures but *explicatures*. As explained earlier, an explicature of an utterance is (roughly) a pragmatically enriched version of its linguistically coded meaning, with ambiguities resolved, references identified, gaps filled, and so on. It can be thought of as the speaker's explicit meaning (as opposed to any distinct implicated meaning).[33] Now, one of the central processes in explicature is *narrowing*, in which the meaning of an expression is narrowed to express a more specific meaning, often an ad hoc, contextually determined one. Noveck and Sperber give the following example:

(15) Henry: Do you want to go on working, or shall we go to the cinema?
Jane: I'm tired. Let's go to the cinema.[34]

'Tired' can be used to express a wide range of physical and mental states from boredom and mild weariness through to outright exhaustion. In the context, however, it is clear that Jane's utterance of 'I'm tired' is relevant only if she means something like 'tired enough to prefer going to the cinema to going on working',

---

[32] Bezuidenhout has made a similar proposal (Bezuidenhout 2002). Unlike Noveck and Sperber, however, she focuses mainly on number terms.

[33] For detailed discussion of explicature, see Carston 2002, 2004a, 2012. As noted earlier, Levinson denies that the distinction between explicature and implicature can be drawn in a rigorous way (Levinson 2000, pp.194–8). However, this does not affect the coherence of the position described in the text. We can agree that narrowing is a real phenomenon, whether or not we think of it as contributing to a distinct process of explicature.

[34] Noveck and Sperber 2012, p.312

and Henry will narrow down the meaning of the term to express that ad hoc concept.

Noveck and Sperber maintain that many cases of supposed scalar implicature are in fact cases of narrowing of this type. As an example, they give the following sentence, which we are to imagine being uttered in the context of a discussion of scientific literacy in America:

(16) Most Americans are creationists and some even believe that the Earth is flat.[35]

Here the hearer will narrow down the basic, semantically coded meaning of 'some' (which Noveck and Sperber take to be 'at least two and possibly all') in the search for a relevant interpretation (that is, one that has contextually useful, easily processed implications). Given the context, it would obviously not be a useful contribution to utter (16) if there were only two Americans who believe the Earth is flat. The utterance is contextually relevant only if a significant number is meant. Moreover, it is common knowledge that not all Americans believe the Earth is flat, so that information has little value, and the contrast with 'most' makes it clear that the speaker means a smaller number than the number of creationists. In this way, the meaning of 'some' is narrowed down at both ends to mean something like 'a number large enough to be relevant to the discussion, but smaller than the number of creationists'.

In cases like this, 'some' is interpreted as having a meaning that is narrowed down *at both ends of the scale*. This narrowed-down meaning will, of course, *entail* the more limited top-end narrowing that the Q-principle would have produced, but it is a much richer and more contextually useful one. Noveck and Sperber give other examples. If Henry is preparing dinner and Jane tells him 'Some of the guests are arriving', a vague reading of 'some' as 'more than one and less than all' will be sufficient to render her utterance optimally relevant (having various contextual implications about what Henry should do next). Similar points, Noveck and Sperber note, apply to other scalar terms. For example, 'possible' may

---

[35] Noveck and Sperber 2012, p.313.

be narrowed down to indicate a modest probability, excluding both certainty *and* tiny probability.[36]

This view provides an alternative, and perhaps more plausible, analysis of some of the cases discussed in the previous subsection. For example, the implicatures in examples (6) and (7), which I suggested would generate implicatures based on nonce Hirschberg scales, may be better thought of as yielding interpretations involving contextually narrowed meanings of 'some' and 'sometimes'. Similarly, the 'tact' implicatures I discussed in the previous subsection might be re-interpreted as resulting from narrowing-down of the key concepts rather than an application of a general Tact principle. In (12) and (13), for example, 'possible' is narrowed to mean 'highly probable' and 'some' to 'many and even all', since these are the meanings from which most contextual implications can be drawn (concerning what actions the hearer should take).

Noveck and Sperber do not claim that we never draw scalar implicatures. They hold that where there is an implicit or explicit question as to whether a stronger term applies, we typically do draw one. If Henry had explicitly asked Jane whether all the guests had arrived, then her utterance would have prompted him to derive the implicature 'not all' (Noveck and Sperber 2012, pp.314–5). (Example (3) in section 2.1 above would be another example of what Noveck and Sperber would regard as a genuine scalar implicature.) But they argue that such cases are much rarer than neo-Griceans believe:

> From the point of view of relevance theory, then, the classical neo-Gricean theory of scalar implicatures can be seen as a mistaken generalisation of the relatively rare case where a weaker claim genuinely implicates the denial of a stronger claim which is under consideration in

---

[36] Noveck and Sperber also give an example in which the meaning of 'some' is *broadened*. Henry has agreed to go and pick up dessert as soon as the dinner guests start arriving, and Jane calls to him 'Some of the guests are arriving'. Here the relevance of Jane's utterance does not depend on how many guests are arriving, and Henry will understand 'some' as compatible with any number of guests arriving, from one to all — which is a broadening of the basic meaning of 'some', as Noveck and Sperber understand it (Noveck and Sperber 2012, pp.313–4).

> the context, to the much more common case where the denotation of an expression is narrowed to exclude marginal or limiting instances with untypical implications. (Noveck and Sperber 2012, p.315)

If this analysis is correct, then it is problematic for Levinson. In cases where considerations of relevance narrow down the meaning of scalar terms, automatic application of the Q-principle will be at best redundant, slowing down the comprehension process rather than speeding it up. If such cases are common, it is hard to see why there would have been pressure for the Q-principle to be applied by default.[37]

It might be suggested that the cases Noveck and Sperber highlight should be regarded as falling under the I-principle, which tells us to enrich interpretations of utterances in the light of background knowledge. I don't think this is a promising suggestion, however, since in these cases the enrichment would depend heavily on contextual factors that have no role in a process of default interpretation (see section 4 below for more on this). Moreover, it is unclear why the Q-principle would not be applied in these cases, and if it is, then, on Levinson's account, the implicatures it produces should override ones produced by the I-principle.

---

[37] Levinson does allow that considerations of relevance (in the everyday sense) may affect how a scalar term is interpreted. He gives the following example:

> A: Is there any evidence against them?
>
> B: Some of their identity documents are forgeries.
>
> (Levinson 2000, p.51)

Here he argues, 'some' is not interpreted as meaning 'not all', since the stronger claim is irrelevant to the speaker's communicative goal (establishing that there is evidence against the people in question). However, on his view this does not mean that the implicature is not derived, but only that is derived and then cancelled.

4. Assessing the I- and M-principles

This section looks at Levinson's other two principles, the I-principle and the M-principle. Again, I shall argue that it is doubtful that these principles support the existence of a level of utterance-type meaning.

4.1 Stereotypes and defaults

The I-principle tells hearers to enrich the content of utterances by drawing on knowledge of relevant stereotypes. Levinson notes that this is a powerful heuristic, which 'allows an interpreter to bring all sorts of background knowledge about a domain to bear on a rich interpretation of a minimal description.'

When introducing this principle, I noted a possible objection. Since the principle tells us to draw on background knowledge to interpret an utterance, it does not look like one that yields default interpretations, associated with utterance types. I suggested that Levinson would reply that the relevant background knowledge can be applied without considering the context of the utterance in question. This is a plausible reply when an expression reliably evokes a single stereotype. Levinson's examples are, arguably, of this kind: 'secretary' is interpreted as 'female secretary', 'road' as 'hard-surfaced road' (though the former might be considered problematic). But, as Anne Bezuidenhout points out, the same expression can evoke different stereotypes in different contexts (Bezuidenhout 2002). As an example, she takes the utterance 'Susan turned the key *and* the engine started.' (an example used by Levinson; 2000, p.117). Here, the I-principle tells us to enrich the interpretation of 'and' by drawing on background knowledge (conjunction buttressing). But (as Levinson himself notes; 2000, p.117), several different enrichments are possible, corresponding to temporal sequence, causal sequence, and goal:

(a) GCI: Susan turned the key and then the engine started.

(b) GCI: Susan turned the key and as a result the engine started.

(c) GCI: Susan turned the key with the goal of bringing it about that the engine started.
(Bezuidenhout 2002, p.266).[38]

Now, in any given context, we will pick out one of these as the relevant stereotype. But plainly the words of the utterance itself cannot determine which is the relevant one, since they are the same in all three cases. In order to access the relevant stereotypical information, it seems we must draw on contextual information. As Bezuidenhout puts it:

> Thus hearers will need to rely on the information made accessible in the wider context, such as information from prior discourse context (i.e., the mutual linguistic context), from the mutual physical environment, or from other shared sources of knowledge. (Bezuidenhout 2002, p.266)

As another example, Bezuidenhout gives 'Professor White's book is on the table'. Here, the I-principle tells us to enrich the possessive (via narrowing this time) by treating it as the stereotypical person-book relation. But again there is no single relation of this kind. The book might be one the professor owns, bought, borrowed, wrote, and so on. And which one the hearer chooses will be determined by the wider context. If the conversation is between assistants in a bookshop, the hearer will probably take the speaker to mean the book the professor wrote; if they are librarians processing requests from academics, they will probably take them to mean the one the professor requested, and so on (Bezuidenhout 2002, p.267–8).

Many more examples could be given. And, of course, similar cases will arise with the M-principle. What counts as the relevant *non-stereotypical* reading of an expression will also vary with context. 'Bill caused to car to stop'[39] implicates that

---

[38] And, as Bezuidenhout notes, this does not exhaust the possible enrichments. 'And' can also implicate relations of temporal inclusion ('He went to London and he saw the Queen'; co-occurrence ('She likes to ride her bike and listen to her Walkman'; enabling ('I forgot to hide the cake and the kids ate it'), and more (examples from Bezuidenhout 2002, p.272).

[39] Levinson's example (Levinson 2000, p.39).

Bill didn't stop the car in the usual way but leaves open a wide range of options, from which the hearer will choose, depending, for example, on whether Bill was the driver, a passenger, a bystander, a policeman, and so on. Indeed, the range of possible M-implicatures will be wider than that of I-implicatures since there are many more ways of being non-stereotypical than of being stereotypical.

All this undermines Levinson's claim that I- and M-inferences are default ones, supporting a level of meaning associated with utterance *types*. If the default interpretation is the one that is most easily accessible, then, as Bezuidenhout notes, expressions will have *many* defaults, varying with context (Bezuidenhout 2002, p.272). Bezuidenhout concludes that Levinson faces a trilemma. If the I-principle produces multiple interpretations of the same expression in every context, then it does not serve the function of speeding up language processing. If it produces different interpretations of an expression in different contexts, then it is not part of a system of default interpretation. And if it produces the same interpretation in every context, then it will often hinder processing, since in many cases this interpretation will have to be overridden and corrected (Bezuidenhout 2002, p.274).

One way of resolving this would be to adopt a weak neo-Gricean view. We might say that hearers employ a general principle which tells them to read unmarked expressions in a stereotypical way (and marked ones in a non-stereotypical way), but that *which* of the many available stereotypes (or alternatives) they settle on will be determined by contextual factors. On this view, the I- and M- principles would guide interpretation, but would not yield default interpretations.

4.2 A deeper problem

There may be a deeper general problem with the I-principle. Levinson holds that the function of the Q-, I-, and M-principles is to overcome the bottleneck in human communication caused by the relatively slow speed of our articulatory processes. But he does not, of course, think that these principles exhaust our pragmatic competence; he assumes that they supplement a system of context-driven pragmatic processing, which processes PCIs and utterance-token meaning

generally.[40] The three principles speed up communication by providing automatic enrichments of certain types of utterance, pre-empting or reducing the need for context-driven pragmatic processing.

Now, it is crucial to this function that they are *formal* principles. Each is triggered by the presence of some formal property (for example, a certain expression type), which can be detected at an early stage of processing. The principle then specifies a formal procedure that can be applied to enrich the content of the utterance in reliable ways. In the case of the Q-principle, the triggering condition is the presence of a lower-ranked expression from a suitable scale, such as 'some', 'possibly', 'may', and the procedure is the replacement of the weaker expression with the negation of a stronger one from the same scale. As we have seen, there are some problems for the Q-principle, but in outline at least it looks like a feasible and effective strategy of default enrichment.

In the case of the I-principle, however, the picture is less clear. The triggering condition here is the presence of an unmarked expression — that is, one that is simple, brief, and familiar. This in itself is problematic, since this condition is the default one: people generally use unmarked expressions, unless they have some reason not to. Given this, it would seem more cost-effective in processing terms to look out for situations in which the condition *doesn't* hold than for ones in which it does. Second, the procedure to be applied is simply an instruction to draw on background knowledge to interpret the expression in the standard way. But this is, presumably, what would have happened anyway, thanks to the context-driven pragmatic processes. In effect, the I-principle says to check if an utterance is

---

[40] He writes, for example:

> In the composite theory of meaning, the theory of GCIs plays just a small role in a general theory of communication. In this regard, GCI theory is not in direct competition with holistic theories like Sperber and Wilson's theory of Relevance, which attempts to reduce all kinds of pragmatic inference to one mega-principle — GCI theory is simply not a general theory of human pragmatic competence. Instead it attempts to account for one relatively small area of pragmatic inference. (Levinson 2000, pp. 21–2)

linguistically normal, and if it is, to process it in the normal way. But this does not pre-empt or reduce pragmatic processing; it is simply permission for it to go ahead, and explicitly applying it would be more likely to delay the process than speed it up. What is the point of checking if a condition holds, unless you're going to do something differently if it does? In short, the I-principle appears to be simply redundant, even from a weak neo-Gricean perspective.[41]

There is a related worry about the M-principle. Here the trigger is the presence of a marked (prolix, unusual) expression, and the procedure is to look for a nonstandard interpretation. This may seem more effective. When a marked expression is detected, we skip the usual processing and jump straight to a nonstandard interpretation. There is a problem, however. How can we tell what a nonstandard interpretation might be until we know what the standard one is? We cannot set aside the standard interpretation until we have identified it. It seems that in order to execute the M-principle, we (or rather, our cognitive systems) will have to let the normal pragmatic processes run until they reach the standard interpretation, *and then let them run further*, looking for alternative, less obvious interpretations.

This does not, however, mean that the M-principle is redundant. It does not pre-empt or reduce context-driven pragmatic processing and does not yield default interpretations. But it does guide how pragmatic processing is conducted and when it terminates. From a weak neo-Gricean perspective, the M-principle may still have a role to play, even if the I-principle does not.

5. Experimental evidence

Neo-Gricean theories were developed by drawing on linguistic intuitions and analyses rather than by experiment. However, since they involve, or at least imply,

---

[41] It might be suggested that processing costs could be reduced by using a single detector, sensitive to only marked expressions, to implement both the I- and M- principles. If the detector is triggered, the M-principle procedure is executed, if it is not triggered, then the I-principle procedure is. However, if executing the I-principle simply involves letting normal pragmatic processing run, this would in effect eliminate the I-principle, since when the detector isn't triggered, nothing different happens.

claims about the mental processes involved in implicature recovery, they are open to experimental testing, and in recent years relevant work has been done in the growing field of *experimental pragmatics* (Noveck and Reboul 2008; Noveck and Sperber 2004). In particular, there are many experimental studies of scalar implicature, designed in part to test whether neo-Griceanism or relevance theory gives a better account of it. In this section I shall briefly survey some of this work and assess its findings.

5.1 Reaction-time studies

A key difference between neo-Griceanism and relevance theory concerns the order in which interpretations of scalar terms are processed. According to Levinson, scalar inferences are made automatically by default, and processing a literal interpretation of a scalar term will involve cancelling the scalar inference. According to relevance theory, by contrast, the initial interpretation is typically the linguistically coded one, and pragmatically enriched interpretations are derived only if needed to meet current expectations of relevance. Thus, in the case of 'some', for example, Levinson's view predicts that the pragmatic meaning 'some but not all' is derived first, whereas relevance theory predicts that the basic meaning 'some and possibly all' is. (The latter meaning is sometimes referred to as the *logical* one, since it corresponds to the existential quantifier of predicate logic). Thus, if Levinson is right, pragmatic interpretations of 'some' should take less time to process than the logical one, and if relevance theory is right, the opposite should be the case.

In a pioneering study (conducted in French), Lewis Bott and Ira Noveck sought to test these predictions (Bott and Noveck 2004). They focused on what they call 'underinformative' sentences, such as 'Some giraffes have long necks', which make a claim that is true on a logical reading of 'some' but false on a pragmatic one (such sentences are said to be *pragmatically infelicitous*). Bott and Noveck used a sentence verification task, in which participants were presented with sentences of the form 'Some/All F are G' and asked to classify each as true or false. A sixth of the sentences were underinformative 'some' sentences, the rest

were control sentences that were straightforwardly true or false.[42] There were two sessions. In one, the participants were told to treat 'some' logically, as meaning 'some and possibly all'; in the other they were told to treat it pragmatically, as meaning 'some but not all'. Bott and Noveck reasoned that if underinformative sentences generate scalar inferences by default, then participants should take longer to respond to such sentences when told to treat them logically than when told to treat them pragmatically, since in the former case the default pragmatic inference would have to be cancelled before the logical reading could be derived. In fact, the opposite happened. Participants responded to underinformative sentences more quickly in the logical condition than the pragmatic one, taking around 800 ms in the former and nearly 1400 ms in the latter. (They also responded more quickly to control sentences in the pragmatic condition, though the difference was not as great.) Participants also gave fewer incorrect answers when instructed to adopt a logical reading (90% correct as opposed to 60% in the pragmatic condition), suggesting that they found it easier to apply the logical interpretation.[43]

In a variant of the experiment, Bott and Noveck allowed the participants to interpret 'some' as they wished. Those participants who classified the underinformative sentences as true were assumed to have adopted the logical interpretation and those who classified them as false were assumed to have adopted the pragmatic one. Again, there was a significant difference in response time, with

---

[42] The control sentences consisted of equal numbers of true 'some' sentences (for example 'Some mammals are elephants'), false 'some' sentences (for example 'Some elephants are insects'), and three sets of 'all' sentences produced by substituting 'all' for 'some' in underinformative, true, and false 'some' sentences.

[43] A possible weakness in the experiment is that the underinformative sentences called for a positive response in the logical condition and a negative response in the pragmatic one. If participants were quicker to confirm a sentence than to deny it, this might explain the difference in response times. To control for this, Bott and Noveck ran a second experiment in which the underinformative sentences called for the same response in both conditions. (They achieved this by asking participants to assess a second sentence that expressed a true/false verdict on the original one and switching the value of the verdict between the conditions.) The results were in line with those of the first experiment.

those who adopted the logical reading responding more quickly than those who adopted the pragmatic one (2700 ms as opposed to 3300 ms).[44]

In a final variant of the experiment, Bott and Noveck manipulated the time participants were given to respond to the sentences presented to them. There were two experimental conditions, Long and Short. In the Short condition participants were allowed 900 ms to respond, in the Long condition they were allowed 3000 ms. Bott and Noveck found that participants were more likely to classify underinformative sentences as true in the Short condition than in the Long condition (72% 'true' responses in the former, versus 56% in the latter). In other words, forcing subjects to respond more quickly (and thus limiting the cognitive resources available for producing their response) increases the likelihood of their treating 'some' as meaning 'some and possibly all' and reduces the likelihood of their drawing a scalar inference 'not all'. (The claim that the availability of cognitive resources affects implicature processing has been confirmed in another study (Pouscoulous et al. 2007). Pouscoulous et al., showed that by using a simpler task, with fewer distracting factors and more basic terms, implicature processing improved across the board from age 4 to adult.)

Bott and Noveck conclude that their studies provide evidence against the neo-Gricean view that scalar inferences are automatic and default and support for the relevance theory view that scalar implicatures take time and effort to process and are derived only when contextually required. The data do certainly indicate that scalar implicatures are not default interpretations, and they thus pose a problem for neo-Griceanism. However, this does not leave relevance theory as the only option.

First, the data are compatible with weak neo-Griceanism. The data suggest that the Q-principle is not applied automatically, before logical interpretations are processed; but it might be applied later and with more effort, if the context makes the logical reading unsatisfactory. That is, the Q-principle may be responsible for scalar implicatures if they are derived, even though they are not derived by default.

---

[44] Another study found an even greater difference, with pragmatic responders taking nearly twice as long as logical ones; see Noveck and Posada 2003.

Second, the data are also compatible with convention theory (or rather its cognitive counterpart). As noted earlier, convention theory treats literal meanings as more basic than generalized ('sentence') implicatures, since it holds that they depend on first-order semantic conventions rather than second-order ones. Moreover, convention theory, I suggest, predicts (at least tentatively) that scalar implicatures will require more time and effort to process than literal interpretations. According to convention theory, deriving literal interpretations involves applying first-order semantic rules only, whereas deriving sentence implicatures involves applying *both* first-order *and* second-order semantic rules. (Second-order semantic rules are rules for expressing further meanings by using sentences with their basic first-order meaning, so a second-order rule cannot be applied until the relevant first-order meaning has been processed. Otherwise, we would be dealing with an idiom rather than an implicature.) This suggests that sentence implicatures should take more time and effort to process than literal meanings. This is only a tentative prediction, of course; to make firm predictions we would need a theory of how knowledge of semantic conventions is stored and accessed. But it is a plausible initial one. Prima facie, then, convention theory fits the experimental data quite well.[45]

---

[45] Other methods are also being used to test theories of scalar implicature. In one of the first studies of its kind, Bezuidenhout and Morris used eye movement monitoring to detect how long participants took to read different regions of a sentence, indicating the different processing demands each region made (Bezuidenhout and Morris 2004). Their aim was to compare Levinson's view (they call it the Default Model, DM) on which scalar terms automatically trigger Q-implicatures, and models such as those discussed in section 3.5 above, where expressions such as 'some' are semantically underspecified and undergo a process of contextually cued pragmatic enrichment (Bezuidenhout and Morris call this the Underspecification Model). Participants were asked to read passages such as the following, in which a 'some' sentence is followed by a sentence explicitly cancelling the supposed 'not all' implicature:

> Some books had colour pictures. In fact all of them did, which is why the teachers liked them.

5.2 Developmental studies

Experimental work has also been done on the development of implicature processing in children. In a pioneering study, Ira Noveck ran a series of experiments to test competence with scalar implicature in French children and adults (Noveck 2001).

In one experiment, participants were presented with underinformative 'some' sentences (such as 'Some elephants have trunks') and control sentences, and asked to say whether they agreed with them. Noveck found that, whereas most adults rejected the underinformative sentences, the majority of the children accepted them (89% of eight-year-olds and 85% of ten-year-olds accepted them, as opposed to 41% of adults), suggesting that most of the children were adopting the logical reading of 'some' and not deriving the implicature 'not all'. (The children correctly evaluated control sentences.) Noveck obtained similar results using the scalar terms 'might' and 'must'. When asked to assess a claim that something *might* be the case (for example, 'There might be a parrot in the box') in a condition in which they knew it *must* be the case, children were much more likely than adults to accept the sentence as true (such sentences were accepted by 80% of seven-year olds, 69% of nine-year-olds, and 35% of adults). Again, this indicates that children tend to adopt a logical interpretation of the modal term, treating 'might' as meaning

---

Bezuidenhout and Morris reasoned that if participants don't make the scalar inference by default when reading the first sentence, but simply start searching for the most contextually appropriate enrichment of 'some' (in line with the UM), then they should spend more time on the word 'all', since it is a strong clue that 'some and possibly all' is the appropriate enrichment. On the other hand, if participants automatically make the scalar implicature to 'not all' (as the DM model predicts), then they will not be pulled up by 'all' since they already have an interpretation of 'some', and it is not until they reach 'them did' that it becomes clear that this interpretation is wrong ('all' might have governed some other predicate). They should thus spend more time processing the words 'them did', which indicate the need for reinterpretation. The results favoured UM rather than DM. In comparison with control sentences in which (for example) 'The books' was substituted for 'Some books', participants spent more time processing 'all' and actually spent less time processing 'them did'.

'possibly and perhaps necessarily', whereas most adults adopt a pragmatic reading, taking the affirmation of possibility to implicate the denial of necessity (Noveck 2001). Noveck concluded that logical interpretations of scalar terms are developmentally primary and that children are, in a sense, *more logical than adults*.

Noveck's findings have been replicated by other researchers (see, for example, Guasti et al. 2005, Experiment 1; Papafragou and Musolino 2003, Experiment 1; Pouscoulous et al. 2007, Experiment 1). Children, it seems, do not spontaneously make scalar inferences. There is evidence, however, that they do have the *ability* to make them, given suitable prompting. After confirming five-year-olds' apparent lack of sensitivity to scalar implicature, Papafragou and Musolino went on to see if they could improve the children's performance by training them to detect pragmatic infelicity. They prepared the children by telling them stories in which a character said 'silly things', which were true but inappropriate (for example, describing a dog as 'a little animal with four legs') and asking how the character might 'say it better'. They also changed the experimental task itself (which involved assessing descriptions of acted-out stories) to make it clear that it was relevant to know whether or not the stronger statements were true.[46] The result was that a much higher proportion of the children rejected underinformative 'some' statements (52.5% of five-year-olds as opposed to only 12.5% in the previous experiment).[47] Moreover, the children who rejected them justified their answer by pointing out that the stronger term was applicable (Papafragou and Musolino 2003).

---

[46] For example, the children would hear about a character Mickey, who had been challenged to put all his hoops round a pole, and, after trying hard, had succeeded. They would then hear Minnie respond to a question about how Mickey had done by saying 'Mickey put some of his hoops round the pole'. The children would then be asked if Minnie had answered well (Papafragou and Musolino 2003, p.271).

[47] The children were also tested on the scales <*finish*, *start*> and <*three*, *two*>, and showed similar increases in pragmatic responding (47.5% vs 10% on <*finish*, *start*>, and 90% vs 65% on <*three*, *two*>. The experiments were conducted with Greek-speaking children.

Other studies have confirmed this. Feeney et al. found that in pragmatically rich contexts (using storyboards and photographs to tell a story and asking participants to assess claims made by one of the characters) only 21% of seven-to-eight-year-olds adopted the logical reading of 'some', as opposed to 57% on a simple sentence verification task (Feeney et al. 2004, Experiment 2). Similarly, Guasti et al. found that in a realistic conversational setting where all the relevant evidence was easily accessible, seven-year-olds derived scalar implicatures at adult levels (Guasti et al. 2005, Experiment 4). Guasti et al. note, however, that the same does not go for younger children. In tests, only half of five-year-olds rejected underinformative statements, even when the statements were presented in a natural way (although the ones that did so, did so consistently) (Chierchia et al. 2001; Papafragou and Musolino 2003). Guasti et al. suggest that at that age some children simply lack the knowledge or ability to derive scalar implicatures, either because the weaker expression does not activate the contrasting stronger one or because the inference from the affirmation of the former to the denial of the latter isn't made (Guasti et al. 2005, p.694).[48]

A possible weakness in some of the experiments reviewed is that they required young children to make metalinguistic judgements (judgements about how well a character had described a situation). These tasks may have been too demanding for younger children, hiding their pragmatic competence. In an ingenious experiment, Yi Ting Huang and Jesse Snedeker sought to get round this problem by using pictures and eye-tracking (Huang and Snedeker 2009). They presented five-year-old children with a range of pictures showing four characters, two boys and two girls, each of whom had a number of items, either socks or soccer balls (but not both). While looking at one of these pictures, the children then heard an instruction of the following form:

---

[48] However, in another study where the task was naturalistic and informational demands clear, children of four-to-five years made scalar inferences at a high level. This extended to particularized scalar implicatures, dependent on nonce scales. For example, if a character was asked whether it had wrapped two presents and replied that it had wrapped one of them, 90% of children detected the implicature that it had not wrapped the other (Papafragou and Tantalou 2004).

Point to the girl/boy that has some/all/two/three of the socks/soccer balls.

Their eye movements were recorded as they listened to the instruction and looked for the intended target.

The pictures were designed in such a way that when 'all', 'two', or 'three' were used in the quantifier position, the identity of the target could be inferred from gender and quantifier alone, and the children tended to look to the target, even before processing the final words ('socks' or 'soccer balls'). When 'some' was used, however, the identity of the target remained uncertain *unless* 'some' was interpreted pragmatically (for example, the relevant options might be a girl with two socks and a girl with *all* the balls). The final words of the instruction then resolved the ambiguity in favour of the pragmatic reading. Huang and Snedeker reasoned that if children made the scalar inference when they processed 'some', then they would look to the correct target before the end of the sentence. In fact, they did not, but delayed looking to the target until they heard the disambiguating words at the end, suggesting that they had not derived the implicature.

Variants of the experiment confirmed this. When the pictures were adjusted so that 'some' identified the correct target whichever reading was adopted (when, for example, the relevant options were a girl with a subset of the socks or a girl with no socks at all), the children looked to the correct target before the end of the sentence. In a final version of the experiment, 'some' was ambiguous, and the pragmatic reading of it now indicated the *wrong* target (for example, a girl with a subset of the balls, when in fact the correct target was a girl with *all* the socks). Huang and Snedeker predicted that if children were drawing the scalar inference, they would take longer to look to the correct target at the end, since they would have to correct an original misidentification. In fact, the adjustment made no difference; the children looked to the correct target just as quickly when it was inconsistent with the scalar inference as when it was not. In all of the experiments, the results from adult participants showed the opposite tendency, indicating that they were drawing the scalar implicature.

As Huang and Snedeker note, this evidence for children's lack of competence with scalar implicature presents a puzzle (Huang and Snedeker 2009, p.1737). For young children are very good at certain kinds of pragmatic processing, in particular

at learning new words through interpreting speakers' communicative intentions. Why then are they so slow to master scalar implicature? Huang and Snedeker suggest that it is because pragmatic processes play different roles in word learning and scalar implicature. In word learning children infer meanings directly from non-linguistic evidence of the speaker's intentions, such as pointing. The pragmatic process is a top-down one, from intentions to meanings, and it can proceed without any prior semantic processing. In the case of scalar implicature, the process is bottom up. The child must start with an analysis of word meaning and move from that to an implicated meaning — which is a much more demanding task. Huang and Snedeker note that other tasks that are hard for young children, such as interpreting irony and metaphor, are also of this bottom-up kind:

> In each case, pragmatic success requires listeners to calculate an interpretation that builds upon but goes beyond the initial linguistic meaning. These postsemantic processes may be particularly difficult, because they require that some feature of the child's initial analysis be revised. (Huang and Snedeker 2009, p.1737)

To sum up then: The experimental data strongly suggest that (a) children initially adopt a logical reading of scalar terms, (b) pragmatic competence increases with age, and (c) children (of seven years and up at least) can derive scalar implicatures if they are provided with suitable contextual assistance.

These results are not what neo-Gricean theory would predict. If the language comprehension system applies the Q-principle automatically, then children should find the pragmatic reading of scalar terms natural and should not need contextual help to derive scalar implicatures. Instead, it should be the logical reading that they find hard to master, since its application will involve cancelling the default pragmatic one. Levinson might reply that the Q-principle is initially applied in a slow and effortful way and only later becomes automatized, but this does not fit well with his view that it is an evolutionary adaptation designed to alleviate the articulatory bottleneck.[49]

---

[49] Levinson writes:

Relevance theory, on the other hand, predicts the experimental results. If logical interpretations of scalar terms are more accessible than pragmatic ones, and if pragmatic interpretations are derived only when required to satisfy expectations of relevance, then we should expect children to derive scalar implicatures less often than adults. For children are typically less aware of informational demands and opportunities than adults and have more limited cognitive resources, which means that they will have lower expectations of relevance and will find implicatures more costly to process and hence less relevant. Likewise, relevance theory predicts that children will draw more implicatures if the context is adjusted to raise their informational expectations and make implicature derivation easier, as they in fact do.

Again, however, there are other options besides neo-Griceanism and relevance theory, not typically considered in the experimental literature. First, the data are broadly compatible with convention theory. It would not be surprising if children learn the first-order rules that govern literal meaning before the second-order rules that govern sentence implicatures. (As noted earlier, Davis predicts that second-language learners will be slower to learn second-order rules, and it may be that children are slower to acquire them in their first language; Davis 1998, p.159.)[50] However, children who have not mastered the conventions for scalar implicature might still be able to work out individual scalar implicatures in a particularized way, drawing on contextual clues and theorizing about the speaker's intentions. This would explain why children derive more scalar implicatures when tested on

---

> Now it is quite clear that … intelligent agents with the asymmetrical abilities in thinking and speaking I have just elucidated, would find a way around the articulatory bottleneck (just as, as a matter of fact, evolution has). The essential asymmetry is: inference is cheap, articulation expensive, and thus the design requirements are for a system that maximizes inference. (Levinson 2000, p.29)

[50] It should be stressed that Davis himself does not commit to the second claim and does not rule out the possibility that first-language learners can master a language's implicature conventions simultaneously with its literal meaning conventions (personal communication).

statements produced in realistic conversational settings where contextual clues are available.

In this context, it is interesting to note that there is evidence that word choice affects derivation of scalar implicatures in children. Pouscoulous et al. found that French nine-year-olds were more likely to draw scalar implicatures when 'quelques' was used for 'some' instead of 'certains', even though their responses on control problems showed that they understood the meaning of 'certains' (Pouscoulos et al. 2007). (42% adopted the logical reading when 'certains' was used, as against 0% when 'quelques' was; the change made no significant difference to adults' responses.) This result is difficult for neo-Griceans to explain, since the words have the same meaning and should both automatically trigger application of the Q-principle. Pouscoulous et al. suggest that 'certains' is a more complex word semantically, which uses up extra processing resources, leaving fewer free for implicature processing — an explanation that fits well with relevance theory. But convention theorists might offer another explanation, suggesting that the implicature conventions governing the two words are different and that those associated with 'certains' take longer to learn.

Second, the data are compatible with weak neo-Griceanism. The evidence indicates that scalar inferences are not drawn automatically whenever scalar expressions are processed, but this is compatible with the view that they are drawn in a more effortful way, when contextually cued. Indeed, there is evidence that even very young children of three-to-four years can draw scalar inferences from contextually salient orderings of non-linguistic stimuli (Stiller et al. 2011; see also Papafragou and Tantalou 2004).

5.3 Tentative conclusions

Three tentative conclusions can be drawn from the experimental research surveyed. First, the data do not support neo-Griceanism as developed by Levinson. Evidence from reaction-time studies and developmental studies suggests that the default reading of scalar terms is the logical one, and that scalar implicature processing is relatively effortful. Second, the data are compatible with the relevance theory. Third, the data are also compatible with other theories of

implicature recovery such as convention theory and weak forms of neo-Griceanism.[51]

6. Conclusions

This chapter has looked at the Gricean framework from the perspective of implicature recovery and linguistic analysis, focusing in particular on the neo-Gricean case for the existence of a class of generalized implicatures derived by the default application of certain general inferential principles. The survey and discussion has necessarily been selective, but it has been sufficient to raise doubts about the neo-Gricean project, at least in the strong form proposed by Levinson. Our examination of Q- I- and M- principles suggests that supposedly generalized implicatures are in fact much more context-sensitive than neo-Griceans suppose, and that default inferences would often need to be cancelled — slowing down, rather than speeding up, the interpretation process. Moreover, the results of experimental work on scalar implicature are at least prima facie incompatible with neo-Griceanism.

The chapter also briefly introduced some alternative approaches to implicature recovery, including relevance theory, convention theory (in a cognitive form), and what I called *weak neo-Griceanism*. I have argued that each of these alternatives has some advantages over neo-Griceanism, but I have not advocated one of them in particular. (I shall return to this topic briefly in the final chapter and suggest that elements of different alternative approaches might be combined.)

I noted at the beginning of this chapter that neo-Griceanism can be seen as offering a theory of implicature generation, with utterances being understood to possess the implicatures they would be interpreted as having according to the GCI principles. Thus, in assessing neo-Griceanism, we have, in effect, also been assessing this simplified and restricted version of the Gricean framework. If the concerns raised in the course of this chapter are sound, then this assessment must be largely negative. If the GCI principles did typically guide our interpretation of utterances in the way neo-Griceans claim, then it would be plausible to give them

---

[51] For further discussion of the experimental literature on scalar implicature, and exploration of its connections with 'dual-process' theories of reasoning, see Frankish and Kasmirli 2010.

this sort of normative status. We could treat them as regularizations of normal practice, and use them to judge specific cases. However, the examples we have considered suggest that the principles do not play the role claimed for them. Even if they do have some role in interpretation (especially, perhaps, the Q-principle), they are not applied by default, in a context-independent way. Details of context and speaker intention can colour implicature recovery even in supposedly generalized cases. Moreover, in so far as there are general patterns of implicature associated with some expressions, these may be better explained as arising from conventions of use rather than GCI principles. So even this limited, simplified version of the Gricean framework looks unpromising. Of course, we could still give the principles a normative status and try to revise our practice to bring it in line with them, but it is hard to see why we should accept such artificial norms, which do not reflect our actual practice or the psychological processes underlying it.

Neo-Griceanism is a bold and elegant theory, but (like the Gricean framework that inspired it) it is too ambitious. Even apparently generalized implicatures can be messy and context-dependent, and they resist simple codification.