5

Three levels of meaning

STEPHEN C. LEVINSON

1 Introduction

Many a student must have sighed when faced with what might seem the almost medieval casuistry of many of the distinctions in John Lyons' (1977) two-volume handbook, *Semantics*. Ambiguities and unclarities of every kind in our frail metalanguage for semantic analysis are there laid out for all to see; a formidable reef of difficult distinctions – types and tokens, acts and products, uses and mentions, originals and replicas, ambiguities of level, etc. – upon which we are all guaranteed sooner or later to founder. Introducing the type/token distinction in a straightforward manner, he goes on to tease us by showing how identifying different tokens of the same type can require a complex measure of similarity or identity of type, and then, having raised our anxieties, announces that it would be 'unnecessarily pedantic' to identify each such distinction (1977: 13–16).

One such distinction Lyons alludes to throughout the volumes may look particularly pedantic, the distinction between utterance-types and utterance-tokens, coming on top, as it does, of the distinctions between system-sentences and text-sentences, sentence-types and sentence-tokens, utterance-acts and -signals and so on. He himself seems to hint (1977: 570ff.) that the distinction may not be of any great utility (since utterance-tokens are rarely constrained to type, and such types could in any case be given formal definition, for example, in terms of sentence-types or forms).

In this chapter I want to suggest that this distinction between utterance-type meaning and utterance-token meaning, or something rather like it, may indeed prove to be an important division in levels of meaning. In finding utility in one of those obscure and seemingly pedantic distinctions, we can be thankful not only to John Lyons but to those generations of scholars in the western tradition, whose work Lyons has so usefully synthesised, who have laboured to hone these fundamental tools of semantic analysis.[1]

2 Levels of meaning

It has long been observed that we need a basic distinction between sentence-meaning and utterance-meaning, where sentence-meaning is understood as the overall meaning composed from the meanings of all the constituents together with the meaning of the constructions in which they occur, while utterance-meaning refers to the import of, say, the very same sentence when uttered in a particular context. (Utterances are thus often treated as pairings of sentences and contexts, namely the contexts in which they occur.) Thus a sentence with deictic elements like *I am sixty-three today* will clearly have different interpretations depending on who says it when, and mismatches between sentence-meaning and utterance-meaning are of course exploited in ironies and other tropes.

This observation is the foundation for the distinction between semantics and pragmatics however this is construed theoretically (see Levinson 1983: ch. 1 for a review; Lyons' *Semantics* of course encompassed both of these levels and the interactions between them).[2] That distinction established two fundamental explanatory levels in a theory of meaning, one responsible for the systematic process whereby the meaning of complex expressions can be built out of the meaning of their parts, and another responsible for explaining how the same expressions might have different meanings or interpretations in different contexts. Theoretical developments will tend to push the boundary one way or the other, but the distinction between the two levels, each with its different explanatory principles, seems certain to survive. It was partly the work of the speech act theorists (Austin and Searle in particular), but especially the work of Paul Grice, that opened up the prospect of a systematic pragmatics. Grice (1957) held that ultimately meaning could be reduced to matters of speaker's intentions, to meaning-$_{nn}$; but proximately, he held that meaning is a composite notion (see Grice 1989). He considered that the full import of an utterance could only be captured by distinguishing many different kinds of content – even the coded content (roughly, our sentence-meaning) was divided between 'the said' and 'the conventionally implicated' (and later he added 'the presupposed'), while the inferred content (our utterance-meaning) was divisible between particularised and generalised conversational implicatures and perhaps other kinds of inference altogether. (See Levinson, 1983: ch. 3, for an introductory exposition.)

It is this distinction between generalised and particularised implicatures that is the focus of this chapter. This distinction, I will argue, should force us to recognise not only the two major levels of a theory of meaning,

semantics and pragmatics, but also a major distinction within pragmatics: a distinction between utterance-type meaning and utterance-token meaning. On general grounds of parsimony, this may be resisted; but I think that by recognising this further bifurcation, we will be greatly aided in understanding the relation of grammar to meaning.

Grice's (1975: 56f.; 1989: 37f.) distinction between particularised and generalised conversational implicatures needs a little exposition. A conversational implicature (henceforth 'implicature' for short), it will be recollected, is an inference that derives from what has been said in context taken together with some general background 'maxims of conversation', enjoining veracity, relevance, perspicacity and the provision of just the right amount of information. Because the inferences derive from both the linguistic expressions and these background assumptions, they are always *defeasible* (or cancellable) whenever the assumptions clearly do not hold. Now, Grice observed, *some conversational implicatures seem context-bound*, while others have a very general currency. Consider, for example, how a single utterance-form might suggest fundamentally different propositions (particularised conversational implicatures or PCIs) in two different contexts, while at the same time implicating something else (a generalised conversational implicature or GCI) in both these contexts and many others too:[3]

(1) Two possible contexts for B's utterance(-form):
a. Context 1:
A: "What time is it?"
B: "Some of the guests are already leaving."
PCI: '*It must be late.*'
GCI: 'Not all of the guests are already leaving.'
b. Context 2:
A: "Where's John?"
B: "Some of the guests are already leaving."
PCI: 'Perhaps John is already leaving.'
GCI: 'Not all of the guests are already leaving.'

The inference labelled 'GCI' here is indeed one of very general currency: normally by stating "Some x are G", a speaker will implicate 'Not all x are G'. So general is the inference indeed that it might be mistaken for part of the meaning of *some* in English; but that it is a pragmatic inference is shown by (a) its predictability by general principle or maxim, (b) the semantic compatibility of its overt denial (as in *Some, in fact all, of the guests are already leaving*).

Some commentators (notably Sperber & Wilson, 1987: 748) have

claimed that Grice attributed no real importance to this distinction,[4] but on the contrary the evidence is that he thought GCIs to be the source of many mistakes in the semantic analysis of, for example, the 'logical' connectives in English. Thus he was keen to point out that the inference from *S's saying "p or q"* to 'S doesn't know that p, or that q' is a regularity of interpretation not to be confused with the conventional or coded meaning of the disjunction. It is the regularity of association that makes the confusion so tempting.

In any case, since Grice much work has shown how useful the notion of a generalised conversational implicature is in linguistic analysis, even if it is not often so explicitly distinguished (see e.g. Gazdar & Pullum, 1976; Gazdar, 1979; Atlas & Levinson, 1981; Horn, 1989). Its utility lies precisely in the idea that certain linguistic expressions will tend to be associated with specific pragmatic inferences across a broad range of contexts, so that these associated inferences can be predicted in a systematic way, and play a systematic role in shaping patterns of lexicalisation and grammaticalisation.

The overall picture of a general theory of communication that then emerges is rather different from the standard picture. According to the standard line, there are just two levels to a theory of linguistic communication, a level of sentence-meaning (to be explicated by the theory of grammar in the broad sense) and a level of speaker-meaning (to be explicated by a theory of pragmatics, perhaps centrally employing Grice's notion of meaning-$_{nn}$). Speaker-meaning, or utterance-token-meaning, will be a matter of the actual 'nonce' or once-off inferences made in actual contexts by specific recipients with all of their rich particularities. This view, though parsimonious, is surely inadequate, indeed potentially pernicious, because it underestimates the regularity, recurrence and systematicity of many kinds of pragmatic inferences.

What it omits is a third layer, intermediate between coded meaning and nonce speaker-meaning, what we may call the level of *statement-* or *utterance-type-meaning*. This third layer is a level of systematic pragmatic inference based *not* on direct computations about speaker-intentions, but rather on *general expectations about how language is normally used*. These expectations give rise to presumptions, default inferences, about both content and force; and it is at this level (if at all) that we can sensibly talk about *speech acts, presuppositions, felicity conditions, conversational presequences, preference organisation* and, of especial concern to us, *generalised conversational implicatures*. It is also at this level, naturally, that we can expect the systematicity of inference that might be deeply interconnected to

linguistic structure and meaning, to the extent that it can become problematic to decide which phenomena should be rendered unto grammar and lexicon and which unto pragmatics (witness the long-standing disputes about the semantic or pragmatic status of illocutionary force and presupposition).

The supposition of this third, intermediate layer in a theory of communication is nothing new. Austin (1962), for example, clearly had something of this kind in mind when he proposed the three-way distinction between locutionary, illocutionary and perlocutionary acts; the locutionary level corresponds to the level of sentence-meaning, the illocutionary to our intermediate layer formed of conventions or habits of use, and the perlocutionary to the level of speaker-intentions. Other theorists have energetically tried to defend the notion of a *convention of use* to be distinguished from a *convention of language*; for example, such a distinction seems essential if we are to retain the idea that indirect speech acts are both partially conventional and inferentially motivated (Searle, 1975). Without admitting the existence of such an intermediate layer, how are we to explain the use of routine formulae (like *Good luck*, *Bless you*, *See you later*) which, although meaning what they literally mean, simultaneously perform habitual everyday rituals (Morgan, 1978)? Why is it that I can introduce myself with *My name is Steve*, but not *I was given the name Steve*; that I can express sympathy with you with *I am really sorry* but not conventionally with *That really saddens me*; that I express outrage with *Really!* but not with *In truth!*; that I can say *I am delighted to meet you* but not idiomatically *I am gratified to meet you*; that I can choose a pastry by saying *I would like that one* but not *I would desire that one* and so on? And to every specification of proper usage there tends to be a restriction on interpretation (Levinson, 1992). There is a great body of language lore here, beyond knowledge of grammar and semantics, extensively studied of course by both ethnographers of speaking and students of second-language learning. That two ways of 'saying the same thing' might be unequal in their conversational import, or that one way of saying something might pre-empt another, these are surely not radical doctrines.

The theory of GCIs is not of course a theory of conventional idioms, clichés and formulae; but it is a *generative theory of idiomaticity*, that is to say a set of principles guiding the choice of the right expression to suggest a specific interpretation, and as a corollary, a theory accounting for preferred interpretations. GCI theory offers a systematic account of why, for example, saying "See you on Tuesday" when tomorrow is Tuesday would suggest not seeing you tomorrow, why saying "If you help, I'll finish it"

suggests that otherwise I will not do so, or why saying "Some of my colleagues are competent" would suggest that not all of them are, and so on, matching a 'way of putting things' with a favoured interpretation in each case. The theory thus belongs to the intermediate level of a theory of communication, the level of utterance-type-meaning.

Nevertheless, that intermediate level is constantly under attack by reductionists seeking to assimilate it either to the level of sentence-meaning or to the level of speaker-meaning; thus, for example, in the case of the inferences we are here calling GCIs, many theorists (Kamp, Peters, Barwise and others) have suggested that they should be in effect semanticised, while Sperber and Wilson and some so-called local-pragmatics theorists have presumed that on the contrary they should be assimilated to matters of nonce-inference at the level of speaker-intention.[5] But generalised implicatures are not going to reduce so easily in either direction, for they sit midway, systematically influencing grammar and semantics on the one hand and speaker-meaning on the other. I shall therefore presume that we do indeed need such a three-tiered theory of communication.

This presumption does not presuppose that the distinctions between the middle layer of utterance-type-meaning and the upper and lower levels is in any way cut and dried. Indeed, there is every reason to suppose that matters of utterance-type-meaning will shade into speaker-meaning at the one end and sentence-meaning at the other. This is in part because there is plenty of evidence that language use is the source for grammaticalised patterns, and that there is a diachronic path from speaker-meanings to utterance-type-meanings to sentence-meanings. Thus grey areas at the boundaries do not constitute evidence against the tripartite view, while evidence for it is the existence of preferred interpretations, default presumptions of the kind we shall illustrate in detail below.

3 Overcoming the bottleneck in human communication: Grice's maxims as heuristics

No student of language can fail to be awed by the intricacies and efficiency of human communication, and the underlying capacities that support it: the specialised physiology, the neurological pathways and the learning abilities that support the structural complexities of language, and above all the sheer miracle of the apparent speed and effortlessness whereby communicative intentions are encoded in articulatory gestures and acoustic signals converted into meanings. It may seem a bit like looking a gift horse in the mouth to point to one part of this miraculous process and identify it as a

relatively slow and inefficient process, which acts as a bottleneck in the entire communicative procedure. Still, if we do so, the finger points inevitably to the articulation process itself: we can think faster than we can speak (e.g. we can do other complex things at the same time, including planning speech ahead), and we can easily understand pitch-corrected speech at double speed, or scan a printed page far faster than it can be read aloud. In fact the psycholinguistic evidence seems to suggest that all the other processes in the entire complex chain of production and comprehension systems could run three to four times faster than the normal pace dictated by the articulation process.[6] Those with a technical turn of mind may like to ruminate on the fact that, even making optimistic assumptions, the transmission rate for human speech is still under 100 BAUD.[7]

The articulation bottleneck in human communication raises interesting questions from, as it were, a design perspective. We can see immediately that any trade-off from coded content to inferential meaning may greatly increase the speed of communication: it will pay to say little and infer much, provided of course the inferential content can be recovered (a) reliably, and (b) speedily. Although we may admire the rich monosyllables of husband–wife communication, the process of recovery of nonce speaker-meaning generally guarantees neither speed nor reliability: the process requires computation of indefinitely nested models of the other's train of thought – what the speaker intended the recipient would think the speaker intended, and so on (see Cohen, Morgan & Pollack, 1990). Even these considerations greatly underestimate the problem of the recovery (by the recipient) of speaker-meaning: there is what might be called the *logical* problem of reconstructed reasoning – since a single conclusion can be reached from an infinite series of different sets of premises, how can the recipient reconstruct the Gricean intentions that lay behind the utterance (Levinson, 1995)?

A much simpler solution would be the provision of some general *default heuristics*, frameworks of assumption that can be taken to amplify the coded content of messages in predictable ways unless there is an indication that they do not apply.

Those default heuristics, I will argue, can be identified with Grice's maxims, or at least a version of them. The heuristics have *default* application; that is, they are applied unless there are explicit indications (in the nature of the context or the content of the message) that they should not be. They then invoke and filter further information of two kinds: information about the structure of the world (or, rather, of stereotypical properties of the relevant domain) and metalinguistic knowledge, that is, information

about semantically related expressions. This information, together with the heuristics and the content of the utterance, provide a set of premises yielding inferences that greatly enrich the informational content of the utterance.[8]

Let me exemplify with three such heuristics, which interact in an interesting way. The details are complex and lie beyond the scope of this chapter, and we must therefore treat them in the most informal way. Let us introduce the cast of characters loosely as follows:

(2) *Three heuristics*

Q1: 'What is not said is not the case'
Constrained to expression-alternates; e.g.
If "x is G" is said, and G and F form a contrast set of expressions, then 'x is not F' is implicated.
Characteristics:
metalinguistic (makes reference to contrast sets e.g. {F, G});
negative (e.g. 'x is *not* F').

Q2: 'What is simply described is stereotypically and specifically exemplified'
(a) unmarked expressions warrant rich interpretations to the stereotype;
(b) minimal forms warrant maximal interpretations.
Constraint: only of unmarked, minimal expressions
Characteristics:
not fundamentally metalinguistic;
invokes world-knowledge of stereotypical relations;
positive inference to specific subcase.

M: 'Marked descriptions warn "marked situation"'
Constraint: only of marked, unusual or periphrastic expressions
Characteristics:
metalinguistic (marked compared to unmarked);
the inference is to the *complement* of the inference that would have been induced by the unmarked expression.

The idea is the following: suppose that the speaker and recipient each know that the other will use exactly these heuristics, then there are many things that will not need to be spelt out (i.e. coded in the linguistic expressions). So, for example, under the first heuristic, if I say "The flag is white", I will implicate (and you will understand) 'The flag is only white, not red, white and blue'. Under the second heuristic, if I say "He opened the door", I will suggest that he entered in the normal way, not using a crowbar or dynamite. Under the third heuristic, if I say "He turned the handle and pushed open the door" I will suggest that he opened the door in some non-stereotypical

manner (e.g. with extra force or speed). In each of these cases, the inference is predictable and clear, and the speaker, knowing this, has – other things being equal – committed himself by a turn of phrase to an interpretation that he knows the recipient will make.

Of course these inferences are defeasible. There is no contradiction in saying "She was wearing a white dress. It had beautiful blue lace trim." Nor, when we are trying to decide whether we are looking at a British or a Russian warship does the observation "The flag is white" or "The flag is red" carry the suggestion 'wholly white' or 'wholly red'. And in complex sentences, the inferences in question may be cancelled by other inferences, as in "They're waving a white flag, even if it's stained red with blood". That is the nature of conversational implicature. Nevertheless, the striking fact is that *ceteris paribus* these inferences do go through by default.

These three heuristics each produce large families of defeasible inferences. By combining all three heuristics, and by presuming that both speaker and recipient will mutually expect them to be in operation, we can greatly amplify the content of what we say – thus overcoming the bottleneck provided by speech-encoding.

4 Default inferences under the three heuristics

The heuristic labels Q1, Q2 and M in (2) above are of course allusions to the corresponding Gricean maxims, the first and second maxim of Quantity and the maxim of Manner.[9] Let us take each of these in turn, and spell out how the heuristics work in a little more detail.

First, Q1, the heuristic that relies on contrast sets of expressions: what exactly is the character of these inferences and from which kinds of lexical sets do they arise? It is clear that there are different kinds of cases, and it is a matter for empirical investigation to find what different kinds of contrast set reliably yield inferences of this negative, complementary kind. Much-studied prototypes are the entailment scales, where we may set up, for example, an ordered pair ⟨S, W⟩ where S is the 'strong' member, and W the 'weak' member, such that when S is substituted in an arbitrary declarative sentence it will entail the same sentence with W substituted for S. In these cases, assertion of the W sentence will carry a generalised implicature that the S variant does not hold, as illustrated below.[10]

(3) a. scale of contrastive expressions: ⟨*all, some*⟩
b. S-sentence: '*All of the students were in class.*'
c. W-sentence: '*Some of the students were in class.*'
d. scalar GCI from the assertion of c: 'Not all of the students were in class.'

These are the prototype cases, and there are many important scalar sets in natural-language vocabularies: all the quantifiers including the cardinal numbers, the truth-functional connectives (⟨*and, or*⟩), many gradable properties (e.g. English ⟨*hot, warm*⟩), many kinds of closed sets of morphemes with so-called 'grammatical meaning' (e.g. English ⟨*the, a*⟩), modal adjectives (e.g. ⟨*necessary, possible*⟩) and much else besides. Closely related to the scalar sets, but yielding inferences of slightly different force, are subordinating connectives of various kinds, including, for example, ⟨*since, if*⟩, as illustrated below (following Gazdar, 1979).

(4) a. clausal subordinators: ⟨*since, if*⟩
b. S-sentence: '*Since Ron saw my manuscript, he's a plagiarist.*'
c. W-sentence: '*If Ron saw my manuscript, he's a plagiarist.*'
d. clausal implicature from assertion of c:[11] 'Ron may or may not have seen my manuscript, and he may or may not be a plagiarist.'

Note that we now have Q1 default inferences attached to most of the 'logical' elements of the vocabulary: to the truth-functional connectives, conditionals, modals and quantifiers. If the systematic pragmatics of these crucial areas of the vocabulary were better appreciated by semanticists, semantic analyses might be rather different and simpler – in character.

In addition to these cases, there are many other kinds of contrast sets capable of yielding systematic Q1 inferences. For example, as illustrated above, the colour terms {*red, white, blue, green*, etc.} denote properties that are extensionally compatible; but asserting that something is red implies that it is not also green, etc., on the grounds typical of Q1 inferences – the speaker can be relied upon to provide enough information (see Harnish, 1976). Indeed, it is possible to plunder the rich observations in structural semantics (as e.g. in Lyons, 1977, or Cruse, 1986) about many different kinds of lexical sets, and explore all the kinds of inferences that may be associated with the employment of individual lexemes from these sets. Note, for example, how the assertion of a superordinate in a taxonomy suggests that the speaker does not know (or thinks irrelevant) which subordinate term or hyponym applies:

(5) a. "I saw an animal in the larder."
b. Q1 implicates: 'I don't know whether it was a mouse, a rat, a squirrel or what.'

There are probably many systematic patterns here yet to be properly explored. (See Hirschberg, 1985, Levinson, forthcoming, for more ideas here.) A cursory inspection suggests a novel idea. The kinds of semantic opposition between expressions in different kinds of sets can be very

different in kind, as explored, for example, in Lyons (1977): as a result the Saussurean notion of *valeur* is then decomposed to the point of loss. Yet we should rightly grieve at the premature death by dissection of a fundamental tenet of structuralist thinking. We might attempt resuscitation by suggesting that *valeur* is not at root a semantic concept at all; perhaps the force is pragmatic, and can be attributed to the Q1 inference to the inapplicability of the contrastive alternate.[12]

Let us turn now to the much less well understood heuristic sketched as Q2 in (2) above ('What is simply described is stereotypically and specifically exemplified'). That there is some such heuristic is indubitable. Consider, for example, the English spatial preposition *in*, as in *in the box*, *in the garden*, *in the cup*. Clearly *in* has a wide range of application: to closed containers (boxes), open containers (cups), bounded spaces (gardens), etc. (see Hirschberg, 1985, for illuminating complexities). Yet when I say "The coffee is in the cup", you do not mistake the relationship between the coffee and the cup for the related but distinct relation indicated in "The pencil is in the cup": we expect partial projection for pencils but not for coffee. It would be more than pedantic to spell out "The coffee is entirely within the bowl of the cup" – more than pedantic because by so saying you would implicate by the third heuristic something other than normal coffee-to-cup relations.

Semantic generality, the large range of applicability of individual expressions, is typical of most of our (non-technical) vocabulary; it is what makes our lexicon of learnable size. Hence Barwise & Perry (1983) have dubbed this property the 'efficiency' of language, neglecting to note that the property would be inefficient indeed without the complementary property of pragmatic enrichment. Semantic generality is also typical of grammatical meaning. Take, for example, the fact that the compositional principles that compute the meaning of phrases specify the composite meaning in only the most general fashion. Thus nominal compounds like *bread knife*, *steel knife*, *murder knife*, *army knife* each have presumptive interpretations along different lines: bread knives are not made of bread, but steel knives are made of steel; murder knives are not made for murder, although army knives are of a type made for armies, and so on. Similarly for the possessive in English: the construction *X's Y* merely indicates that *some* relation holds between the two noun phrases, and we resolve the relation by pragmatic inference. Thus the phrases *Jupiter's moons*, *John's ideas*, *Anne's address*, *the building's condition*, *the encyclopedia's editor*, *the year's end*, are each understood to involve different relations (gravitational capture, ideational authorship, postal access, etc.). Note that all these phrases seem to have a default interpretation: *John's pens* will naturally be taken to mean the pens

belonging to John, unless the context (e.g. talk between pen-designers) warrants another less stereotypical interpretation.

What is clear then is that, hearing an utterance, we imagine a specific instantiation, a stereotypical exemplification. But why should such a tendency, perhaps psychological in origin, constitute a heuristic? Are we not confusing the private interaction between our individual knowledge of the world and our understanding of utterances with a theory of communication? The answers lie in the strange power of the reflexive reasoning that Grice introduced in his theory of meaning-$_{nn}$. The speaker, knowing the recipient's interpretation to tend in a particular direction, and knowing that the recipient knows the speaker so knows, can turn a good chance into a certainty: mutual awareness of the interpretation to the stereotype guarantees that this is what the speaker intends. The speaker *designs* his or her utterance accordingly. (The same principle holds, *mutatis mutandis*, for all the heuristics, of course.)

Closely allied to the inference to the stereotype is a class of other inferences to the more specific subcase. Many of these have to do with the maximisation of coherence, the minimisation of postulated entities and the presumptive enrichment of mentioned relations. For example, it has long been noted that conjunction, or in many languages paratactic adjunction, is presumptively enriched to suggest sequential occurrence of events and, further, intention and causality, as illustrated below, where the assertion of (a) will suggest (b), (c) and (d) even in the absence of stereotypical connections between bells and engines:

(6) *Conjunction-buttressing* (Atlas & Levinson, 1981)
  a. "Ann rang the bell and the engine started."
  b. 'Ann rang the bell and *then* the engine started.'
  c. 'Ann rang the bell and *therefore* the engine started.'
  d. 'Ann rang the bell, *thereby intending* the engine to start.'

A similar presumptive strengthening of content is typical of conditionals, as illustrated below, where the assertion of (a) will suggest (b) and thus jointly (c):[13]

(7) *Conditional perfection* (Geiss & Zwicky, 1971)
  a. "If you co-operate, there'll be no trouble."
  b. 'If you don't co-operate, there will be trouble.'
  c. 'If and only if you co-operate, will there be no trouble.'

Negative statements are of course informationally weak: from the assertion that x is not F, one is left in the dark as to whether x is G or H, etc. They are thus ripe for pragmatic enrichments of many kinds (see Horn, 1989), but a

genus that comes under the rubric of the Q2 heuristic includes the many cases where contradictories are routinely 'read as' contraries:

(8) a. "I don't like the new boss."
b. 'I positively dislike the new boss.'
(9) *'Negative-raising'*
a. "I don't believe he will show up."
b. 'I believe he will not show up.'

Another wide class of Q2 inferences involves interpretations that maximise cohesiveness – anaphoric linkages, for example. It is well known that anaphoric linkages are made partly on the basis of encyclopedic knowledge, but there are also clear preference patterns, for example for local (proximal) coreference which can be demonstrated in texts, as illustrated below:

(10) Then the thief$_1$ asks the butler$_2$, say, and the butler$_2$ confirms that. He$_2$ says, 'Yes the superintendent has only just left.'
(from Agatha Christie, *Hercule Poirot's Christmas*)

The general heuristic seems to be: do not postulate more entities than necessary, and link locally by preference.

No doubt rather specific mechanisms are involved in each of these preferences, including the inferences to the stereotype, that we currently do not fully understand; but that there are such preferences – and not just calculations of speaker-meaning – seems rather clear. Gathering them together under the rubric of the Q2 heuristic is not simply a matter of convenience, for the inferences share certain crucial properties. First, they are inferences to *more specific* interpretations, where what is implicated is a subcase, a specific instantiation, of what is said. The inferences are positive and non-metalinguistic in character, unlike Q1 or (as we shall see) M inferences. They are default inferences – not all inferences to the subcase have this character. They are tied to the use of unmarked, 'minimal' or non-prolix, semantically general expressions (or even the absence of them as in parataxis or zero anaphora). Note, for example, the following interpretative contrasts between minimal expressions (italicised in the (i) examples) and more marked expressions (italicised in the paired (ii) examples) that might be thought to paraphrase their content:

(11) a. (i) John pushed the button *and* the motor started.
(ii) John pushed the button. *In addition*, the motor started.
b. (i) The detective came into the room and *he* sat down.
(ii) The detective came into the room and *the man* sat down.
c. (i) I *don't like* garlic.
(ii) I *have no liking* for garlic.

d. (i) The book is *on* the desk.
   (ii) The book is *in contact with the upper surface of the desk*.
e. (i) *John's picture* won critical acclaim.
   (ii) *The picture of John* won critical acclaim.

In addition to these shared properties, Q2 inferences share similar projection properties, to be noted below.

Any kind of inference from a general description to the special subcase clearly must be strictly constrained: we make no inference from the assertion "John drives a small car" to 'John drives a Honda Civic' even if the probabilities are so. Indeed, as discussed above, "I saw an animal in the larder" suggests that I do not know what animal I saw. How then are Q2 inferences constrained? Partly they are constrained with respect to depth; and here the notion of stereotype needs explication – as Putnam pointed out, a stereotype has nothing to do with statistical tendencies, or even with shared veracities. Fierce gorillas, gentle cows, absent-minded professors are stereotypes for which there is little evidence or even shared belief. There is no such stereotype from small cars to Honda Civics, and the speaker knows the addressee knows that the speaker will not presume so. But male surgeons are another matter, and there are many parlour puzzles of the sort "The patient went to see the surgeon. She described the problem to him and she decided at once to operate on him". Inferences to the stereotype are thus not 'generalised' in the sense that they are independent of shared beliefs (as Q1 and M inferences largely are, since they are essentially based on metalinguistic considerations), but they are 'generalised' in the sense that they follow a general principle – restrict the interpretation to what by consensus constitutes the stereotypical, central extensions.

More importantly perhaps, Q2 inferences are constrained by the other heuristics. Any Q1 inference incompatible with a Q2 inference always takes precedence. Any M inference from a marked expression likewise defeats a Q2 inference, in ways that will be explained. The result is that a Q2 inference is induced by a certain kind of expression, especially expressions that are themselves brief and colloquial. Like the following heuristic that bounds it, Q2 is thus iconic: minimal expressions invite stereotypical, rich interpretations.

Finally, we turn to the third heuristic introduced in (2) above ('Marked descriptions warn "marked situation" '), labelled M after Grice's maxim of Manner.[14] Now, we have already seen from examples in (11) that marked or more prolix expressions do not give rise to the Q2 inferences that would have arisen from their unmarked or brief counterparts. In fact, there is a systematic complementarity between unmarked expressions and their

associated Q2 inferences compared to marked expressions and their M inferences.[15] The relevant sense of 'markedness' is very broad, covering formal prolixity, infrequent expressions or those of unusual formation – the M-principle is again iconic: 'non-stereotypical expressions invite interpretations to non-stereotypical extensions'. Take, for example, the following lexical doublets, and the sort of denotation they might suggest in some arbitrary utterance (the symbol + > should be read 'implicates, *ceteris paribus*'):[16]

(12) a. unmarked: *drink* Q2 + > 'alcoholic drink'
marked: *beverage* M + > 'non-alcoholic drink'
b. unmarked: *chair(man)* Q2 + > 'male chair person'
marked: *chairperson* M + > 'female chair person'
c. unmarked: *knife* Q2 + > 'kitchen-type knife'
marked: *cutter* M + > 'not a normal knife'
d. unmarked: *missile* Q2 + > 'rocket with warhead'
marked: *projectile* Q2 + > 'missile other than rocket'
e. unmarked: *letter* Q2 + > 'personal letter'
marked: *missive, dispatch, epistle* + > 'not a personal letter'
f. unmarked: *house* Q2 + > 'normal family house'
marked: *residence* M + > 'grander than normal family house'
g. unmarked: *rare* Q2 + > 'unusual and valuable'
marked: *scarce* M + > 'in short supply'

Similarly for word formation: derivations tend to sort into two classes, the usual, colloquial with a specialised stereotypical extension, and the more unusual or prolix derivation picking up (often now by convention) the complementary interpretation (see Horn, 1989: 273ff., for discussion):

(13) a. unmarked: *informer* Q2 + > 'supplier of information against someone'
marked: *informant* M + > 'supplier of information for someone'
b. unmarked: *unnatural, unscientific* Q2 + > 'and bad'
marked: *non-natural, non-scientific* M + > '(no special evaluation)'
c. unmarked: *imprecise/immoral*, Q2 + > 'the opposite of precise/moral' (i.e. the contrary reading)
marked: *unprecise/non-* or *amoral* M + > 'just not precise or not moral' (i.e. the contradictory reading)

There is also an opposition between simplex lexemes and derived forms which might be thought to have the same meaning and use (e.g. *sad* vs. *unhappy*, or *rude* vs. *impolite*, where the lexicalised form invariably seems to denote a more extreme property; again, see Horn 1989: 279f., also Kiparsky, 1983).

These sorts of pragmatic principles explain how specific kinds of word form may acquire specialisations of meaning: they are principles that may explain historical changes and semantic shifts. By the same token, the synchronic analysis of current lexical content is sometimes of course not so clear.

Rather clearer cases of the Q2 vs. M opposition may therefore be found in periphrastic alternatives to simple lexicalisations. Thus periphrastic modals, causatives and double negations contrast with their simpler counterparts:

(14) a. "John could solve the problem" Q2 + > 'and he did'
b. "John had the ability to solve the problem" M + > 'but he didn't'
(15) a. "James stopped the car" Q2 + > 'in the normal way, by using the foot pedal'
b. "James caused the car to stop" M + > 'in a nonstereotypical way, e.g. by using the hand-brake'
(16) a. "Sue moved the car" Q2 + > 'by driving it'
b. "Sue made the car move" M + > 'e.g. by pushing it'
(17) a. "It's possible he will recover" Q2 + > some definite probability p
b. "It's not impossible that he will recover" M + > some probability less than p
c. "The mail is reliable" Q2 + > 'to degree *n*'
d. "The mail is not unreliable" M + > 'to degree less than *n*'

Repetition and reduplication also serve to deflect interpretation from Q2-directed extensions:

(18) a. "He ate" Q2 + > 'He ate the normal meal.'
b. "He ate and ate" M + > 'He ate more than the normal meal.'

In many languages, reduplication plays an important quasi-derivational role, and it is notable how such reduplications tend to pick out, not the central or prototypical extensions of the unreduplicated form, but their complements, the peripheral or non-stereotypical extensions.[17]

These three heuristics are each responsible for large families of inferences, each of a characteristic type. In certain ways the principles are quite clearly antagonistic: they encourage inferences in opposite directions. Thus whereas Q2 invites inferences to the more specific subcase (along the lines of 'The speaker hasn't said what is obvious'), Q1 forbids the inference to the *more informative interpretation (along the lines 'If the speaker didn't say it*, he didn't mean it'). Similarly, M1 inferences are specified as the complement of Q2 inferences. Contradictory premises would be fatal to any deductive device, and problematic for any inferential system. They must

therefore be blocked at source, or filtered by simple rule or procedure. In fact, both mechanisms seem to be involved. For example, many apparently potential Q1 inferences do not in fact arise because there are strict criteria of both form and content on the sets that give rise to them (Atlas & Levinson, 1981; Levinson, 1987a: 407). In addition, there are simple priority rules of the kind: Q1 and M inferences take precedence over inconsistent Q2 inferences; Q1 inferences take precedence over inconsistent M inferences (Levinson, 1987a: 409). Within each genus there also seem to be priorities: thus clausal Q1 inferences cancel inconsistent scalar Q1 inferences (as noted and formalised in Gazdar, 1979). In short, there is a serious projection problem for generalised implicatures, but fortunately we already have some understanding of how the problem is resolved.[18]

5 Grammar and meaning

The mechanisms reviewed here – a set of three general heuristics that induce default inferences – have completely general application across the vocabulary of a language; but they may yield inferences that are particularly precise, specific and recurrent where small closed sets of lexemes or morphemes yield contrast sets of the right kind to induce Q1 implicatures. Such sets are typical of the 'grammatical' or 'functor' words. For example, as noted above, early Gricean analyses pointed out that the English sentential connectives form a Q1 entailment set ⟨*and, or*⟩, so that an utterance of the form "p or q" will generally implicate 'not p and q', while "p and q", unbounded by a Q1 inference, will (where p and q describe events) Q2 implicate 'p and then q' and so on. Such an analysis allows us to hang on to the simple underlying invariant meaning of the connectives, while explaining all the additional variable readings.

Exactly the same kind of analysis can now be applied to any grammatically closed class of morphemes, and should yield the same kind of harvest: invariant core meanings supplemented by preferred interpretations. Take, for example, the definite/indefinite articles in English. Simple accounts of the kind 'introduce a new referent under description Y with *a Y* and a previously mentioned one under description X with *the X*' or 'Use *the X* to refer to a unique entity, *a Y* to refer to a non-unique entity' run afoul of multitudinous counter-examples, as shown by Hawkins (1978, 1991). For example, it is quite normal to introduce some entities with *the X*:

(19) a. I just met the Mayor.
b. I'm late because I missed the train.

c. There's something wrong with the clutch (in my car).
d. She adores the man she met in Paris.

and indefinite articles can be used for previously introduced referents:

(20) a. *All the members of the jury met for many hours because a single member* was recalcitrant.
b. His arms and legs were damaged in the blast, and in the end he lost *a leg*.

while some unique entities are happily referred to with an indefinite article:

(21) a. England has *a Queen* and Spain *a King*.
b. There is *a dog* in that car.

Hawkins (1991) points out that if we adopt an account in terms of GCIs most of these puzzles evaporate. The articles form a Q1 entailment contrast set ⟨*the*, *a*⟩, such that *the X* entails uniqueness, while *a Y* only implicates non-uniqueness, which may thus evaporate in contexts like those immediately above. Hawkins argues that *the X* conventionally (i.e. non-defeasibly) implicates that there is a mutually salient set in which X is unique, while *a Y* Q2-implicates (i.e. defeasibly) that there is a mutually salient set to which Y belongs (hence "He lost a leg" suggests one of his own).

There are of course many other closed sets of grammatical contrasts, often notorious for their semantic subtleties, that could benefit from a pragmatic analysis of this kind, for example, deictic adverbs or determiners, tense and aspect markers and prepositions.[19]

Further inroads into 'grammatical meaning' may be possible. There has been much speculation about whether a pragmatic analysis might undercut the purely grammatical analysis of anaphoric dependencies typical of modern grammatical theory (Reinhart, 1983; Levinson, 1989, 1991; Huang, 1994). For example, we can think of the opposition between non-reflexive and reflexive pronouns as similar to that between indefinite and definite pronouns: wherever *him* and *himself* can contrast, for example in direct-object position, we have a Q1 scale of the kind ⟨*himself*, *him*⟩ such that use of *him* will implicate 'not himself' – rather than be grammatically stipulated as non-coreferential by Binding Condition B in the Government–Binding framework of Chomsky (1981). The advantage of the pragmatic account is that, since GCIs are only default inferences, it allows the possibility of coreference between subject and a non-reflexive pronominal object in unusual cases (such as *Only Felix$_i$ voted for him$_i$* – see Reinhart, 1983). Given Binding Condition A (which stipulates that reflexives must find their antecedents in certain positions), the other Binding Conditions (Binding Condition B governing the non-coreferential

interpretation of non-reflexive pronouns in certain positions and Binding Condition C stipulating the non-coreferential interpretation of full lexical NPs) are regularities at least partially predicted by the system of heuristics here outlined.[20] In sketch form, consider the following patterns:

(22) a. John$_i$ likes himself$_i$.
(stipulated by Binding Condition A)
b. John$_i$ likes him$_j$.
(Q1 inference from the non-use of *himself* – pattern often attributed to Binding B)
c. John$_i$ told her that he$_i$ would come.
(Q2 inference to coreference, unblocked by Q1 inference since *himself* cannot occur in this position)
d. John$_i$ told her that the man$_j$ would come.
(M inference to an interpretation contrastive with that of unmarked, simpler form in (c); pattern often attributed to Binding C)

This is not the place to pursue this analysis, which can be refined – and of course countered – in various ways (see Levinson, 1987a, 1991). The point to be made here is that even if we decide that in fact these anaphoric patterns are grammaticalised in English, the very possibility of a (perhaps incomplete) pragmatic analysis undercuts Chomsky's view that the patterns in question must be specified by native 'universal grammar' because they are abstract and unpredictable from usage patterns, and thus in effect unlearnable.

The pragmatic point of view seems to be supported by the facts from languages (like many Austronesian ones, Biblical Hebrew, old Germanic languages, some Australian languages) which do not exhibit reflexives at all. In these languages a sentence glossing 'John hit him' may have a reflexive or non-reflexive reading, with the latter the default. The default reading presumably arises (as a Q2 inference) from the stereotypical agentive schema, in which an agent acts on another entity (Farmer & Harnish, 1987). To block this disjoint reading, the pronoun is normally marked by an emphatic particle or affix to indicate by M inference that the complementary interpretation is intended. Elsewhere, outside the clause, the pronoun tends to pick up coreferential readings as in English. Thus we have in schematic gloss the following pattern:

(23) a. 'John$_i$ likes him$_j$.'
Q2 inference to stereotypical action
b. 'John$_i$ likes him EMPHATIC$_i$.'
M inference to complement of (a)

c. 'John$_i$ told her that he$_i$ would come.'
Q2 inference to minimal domain of discourse
d. 'John$_i$ told her that the man$_2$ would come.'
M inference to complement of (c)

This pattern in fact suggests a general diachronic source for true reflexives in marked, emphatic forms of a normally disjoint pronoun. Certainly the history of English reflexives has normally been analysed this way (see Visser, 1963: 420–39), and in languages with continuous and ancient written traditions, like Japanese, Tamil and Chinese, there seems to be evidence in the same direction (see Faltz, 1985, and references in Levinson, 1991), while a swifter development of the same kind can perhaps be observed in creoles (Carden & Stewart, 1987, 1988).

Again, we cannot pursue these issues here (see Levinson, 1991), but the general point is that there are languages which have no reflexives, and a corresponding freedom of anaphoric interpretation (in part because they lack the strong Q1 inferences that play off the opposition between reflexive and non-reflexive pronoun). Such languages would seem to be anomalous to the Government–Binding framework, but the patterns of interpretation seem rather well predicted by our heuristics or something like them (see Huang (1991, 1994) for a developed account along similar lines of the flexible patterns of interpretation in Chinese).

In short, patterns of preferred interpretation may play an important role in the relation between grammar and meaning: grammatically or lexically stipulated meanings tend to generate a set of further default interpretations from the use of related but distinct forms. These in turn can become conventionalised or grammatically stipulated, yielding yet further default inferences. Given these diachronic tendencies, the analyst may easily mistake a default inference for a lexically or grammatically stipulated meaning, and of course vice versa (mistaking a conventionalised ex-inference for a live one).

6 Conclusion

This essay has argued for a fresh perspective on linguistic communication, where more attention is given to preferred ways of 'putting things', or the use of favoured constructions for conveying specific messages.[21] Instead of a bifurcation between grammatically and lexically specified meaning and nonce speaker-meaning, we need to develop a three-tiered theory of communication in which utterance-type meaning has a special place. The

theory of utterance-type meaning should be a theory of default interpretation. This level of meaning may exhibit some relatively tight universal constraints, because (or so I have suggested) it is based on a set of heuristics that are designed to overcome an intrinsic bottleneck in the speed of communication, our slow articulation rate. Because this level of meaning sits midway between grammar and lexicon on the one hand and speaker-meaning on the other, most analysts attempt to reduce it in one or the other direction. This is a mistake, because it is a level with distinct properties – default, defeasible inferences based on the comparison of alternative linguistic expressions and on the presumption of stereotypical situations, which interact in specific ways. In addition, because these systematic mechanisms are so closely related to grammatical and lexical processes, they constrain them and, over the course of language history, feed them.[22] It is thus quite unlikely that we will have an adequate synchronic or diachronic theory of grammar and lexicon until we have a much deeper understanding of the level of utterance-type meaning.

NOTES

1 This paper is a sketch of issues treated in more depth in Levinson (forthcoming: ch. 1). A prior, spirited defence of the idea of three levels of meaning may be found in Atlas (1989: 3–4 and *passim*).

2 Lyons thus uses the term 'semantics' in a pretheoretical way to denote the full range of linguistic meaning; I will use the term in the narrower sense, opposed to pragmatics. I will, however, continue to use the term 'meaning' in the wide sense, not restricting it to coded morpheme- or sentence-meaning. For the larger field, the study of meaning in this wide sense, I will use the phrases 'the theory of meaning' or, where the wide range might not be clear, 'the theory of (linguistic) communication'.

3 Let us adopt the typographical conventions that utterances (or rather utterance-types) are indicated by double quotes, interpretations or glosses by single quotes, linguistic expressions or sentences by italics. We will also use the symbol +> for '(generally) conversationally implicates', so that "Some boys are naughty" Q2 +> 'not all boys are naughty' is read as 'The uttering of the sentence *Some boys are naughty* will by default inference under the Q2 heuristic have the additional interpretation "not all boys are naughty".'

4 Sperber & Wilson (1987: 748) wish to abolish GCIs because their proposed account of how implicatures are calculated cannot accommodate the phenomenon (see Levinson, 1987b, 1989). For other kinds of problems with the distinction, see Hirschberg (1985: 42).

5 For a representative attempt to semanticise the phenomena, see, for example, Barwise (1986); for representative attempts to reduce GCIs to nonce inferences, see, for example, Hobbs (1987), Kempson (1986) and of course Sperber & Wilson (1986).

6 For the comparative speed of pre-articulation vs. articulation processes, see Wheeldon & Levelt (forthcoming). For the ability to parse and comprehend speeded speech, see Mehler *et al.* (1993).

7 The calculation, kindly made for me by Bill Poser, assumes 7 syllables or 17.5 segments per second, and 5.5 bits per phoneme.

8 I will not here discuss the nature of the inference itself. With regard to the nature of implicature generally, there are divergent opinions: Sperber & Wilson (1986) maintain it is deductive, Grice (1973) explicitly likened it to inductive inference, while Atlas & Levinson (1981) suggested that some inferences have an abductive character. GCIs, though, are by hypothesis default inferences, both non-monotonic and presumptive. There is now a large family of formal models for such systems of inference: see, for example, the collection in Ginsberg (1987).

9 The labels Q1, Q2 and M, adopted here in deference to Grice's maxims, refer respectively to the Principles Q, I and M (or Q/M) in Levinson (1987a, 1991). Comparison with Horn's (1989) system will be aided if it is noted that his R is my Q2 (or I), while he conflates my Q1 and M under a single rubric Q. All three of my principles are conflated into one R (or Relevance principle) in Sperber & Wilson's (1986) proposal. The profusion of proposals indicates of course that this is now an active research area.

10 For much further detail see Horn (1972, 1985, 1989), Gazdar (1979), Atlas & Levinson (1981) and Hirschberg (1985). As Gazdar points out, the inferences are epistemically modified, in ways that are crucial to any formalisation, but which we ignore here.

11 That this inference is defeasible, therefore pragmatic, is shown by reasoning of the kind *He has AIDS. If he has AIDS, his wife has too.*

12 See also Clark (1993: ch. 5) for the view that *contrast* is a pragmatic strategy for language learning.

13 I have been loose about distinguishing what is implicated from what is said-and-implicated, although there is no particular problem in doing this. Incidentally, the 'conditional perfection' kind of inference is independent of the indirect illocutionary force of such utterances – promises, threats, predictions but also plain conditional assertions tend all to carry the inference.

14 The reference is particularly to Grice's first and fourth submaxims of Manner: 'avoid obscurity' and 'be brief'.

15 This observation is due to Horn (1985), who points, however, to a long tradition of essentially similar analyses in the study of morphology.

16 For reasons of space I have not spelt out the contexts of use in which the doublets might reasonably be claimed to have the same semantic content; obviously such lexical doublets are likely to have some divergence in use other than those explained by our pragmatic principles here.

17 For example, in Tamil there is a productive reduplication with rule-bound phonological alteration: thus *paittiyam* 'madness' becomes *paittiyam-giytttiyam* 'almost but not quite real insanity'. Moravcsik (1978) gives a partial account of some of these patterns. She points out that there are very different predictions where reduplication is the only signal available for some interpretation (as it is for plurality or repetition in many languages) – M implicatures after all only function by contrast to another simpler way of 'saying the same thing'.

18 Gazdar's (1979) formal system might be adapted to handle many aspects of the observable cancellation properties of the different kinds of inferences. In effect, we would set up an incremental augmentation of the contextual assumptions in a specific order: entailments > Q1-clausal > Q1-scalar > M > Q2, etc., such that inferences are added in that order only if they are consistent with what is already taken for granted. But certain problems remain: there are, for example, constraints beyond consistency.

19 Spatial prepositions in English are an interesting case: we can set up Q1 contrast sets of the kind ⟨*at*, *near*⟩, such that "The train is near the station" suggests 'The train is not (yet) at the station', and so on. Q2 inferences from prepositions like *in* to the relevant stereotypical relations have been illustrated above, while M contrasts like {*on*, *on top of*} are also easy to find.

20 Much more needs to be said of course about c-command constraints on interpretation; see Levinson (1987) and discussion there of Reinhart's (1983) proposals.

21 I borrow here the emphasis on 'favoured constructions' from John Haviland.

22 For many insights into pragmatic constraints on lexicalisation, and diachronic processes, see Gazdar & Pullum (1976), McCawley (1978) and Horn (1985, 1989).

REFERENCES

Atlas, J.D. (1989) *Philosophy without Ambiguity: a Logico-linguistic Essay*, Oxford: Clarendon Press.

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

Austin, J.L. (1962) *How to Do Things with Words*, Oxford: Clarendon Press.

Barwise, J. (1986) Conditionals and conditional information. In E. Traugott. A. Ter Meulen, J. Reilly & C. Ferguson (eds.) *On Conditionals*, Cambridge University Press, 21–54.

Barwise, J. & J. Perry (1983) *Situations and Attitudes*, Cambridge, MA: MIT Press.

Carden, G. & W. Stewart (1987) Mauritian creole reflexives – an alternative historical scenario. MS, University of British Columbia.

(1987) Binding theory, bioprogram and creolization: evidence from Haitian creole. *Journal of Pidgin and Creole Linguistics* 3: 1–67.

Chomsky, N. (1981) *Lectures on Government and Binding*, Dordrecht: Foris.
Clark, E. (1993) *The Lexicon in Acquisition*, Cambridge University Press.
Cohen, L.J. (1971) The logical particles of natural language. In Y. Bar-Hillel (ed.), *Pragmatics of Natural Language*, Dordrecht: Reidel, 50–68.
Cohen, P., J. Morgan & M. Pollack (1990) *Intentions in Communication*, Cambridge, MA: MIT Press.
Cole, P. (1978) *Syntax and Semantics*, vol. IX: *Pragmatics*, New York: Academic Press.
Cole, P. & J. Morgan (1975) *Syntax and Semantics*, vol. III: *Speech Acts*, New York: Academic Press.
Cruse, D.A. (1986) *Lexical Semantics*, Cambridge University Press.
Faltz, L.M. (1985) *Reflexivization: a Study in Universal Grammar*, New York: Garland.
Farmer, A. & M. Harnish (1987) Communicative reference with pronouns. In M. Papi & J. Verschueren (eds.) *The Pragmatic Perspective*, Amsterdam: Benjamins, 000–00.
Gazdar, G. (1979) *Pragmatics*, New York: Academic Press.
Gazdar, G. & G. Pullum (1976) Truth-functional connectives in natural language. In *Papers from the thirteenth regional meeting of the Chicago Linguistic Society*, Chicago Linguistic Society, 137–46.
Geiss, M. & A. Zwicky (1971) On invited inferences. *Linguistic Inquiry* 2: 561–5.
Ginsberg, M.L. (ed.) (1987) *Readings in Non-monotonic Reasoning*. Los Altos: Morgan Kaufman.
Grice, H.P. (1957) Meaning. *Philosophical Review* 67: 377–88.
(1973) Probability, defeasibility and mood operators. MS, paper delivered to the Texas Conference on performatives, presuppositions and implicatures, 1973.
(1975) Logic and conversation. In P. Cole & J. Morgan (eds.) *Syntax and Semantics*, vol. III: *Speech Acts*, New York: Academic Press, 41–58.
(1989) *Studies in the Way of Words*, Cambridge, MA: Harvard University Press.
Harnish, R. (1976) Logical form and implicature. In T. Bever, J.J. Katz, & T. Langendoen (eds.) *An Integrated Theory of Linguistic Ability*, New York: Crowell, 313–92.
Hawkins, J. (1978) *Definiteness and Indefiniteness: a Study in Reference and Grammaticality Prediction*, London: Croom Helm.
(1991) On (in)definite articles: implicatures and (un)grammaticality prediction. *Journal of Linguistics* 27: 405–42.
Hirschberg, J. (1985) *A Theory of Scalar Implicature*, Moore School of Electrical Engineering, University of Pennsylvania, MS-CIS-85-56.
Hobbs, J. (1987) *Implicature and Definite Reference*, Stanford: CSLI Report 87–99.
Horn, L. (1972) *On the Semantic Properties of the Logical Operators in English*, Indiana University Linguistics Club mimeo.
(1985) Toward a new taxonomy for pragmatic inference: Q- and R-based implicature. In D. Schiffrin (ed.) *Meaning, Form and Use in Context*, Washington: Georgetown University Press, 11–42.

(1989) *A Natural History of Negation*, Chicago University Press.

Huang, Y. (1991) A neo-Gricean pragmatic theory of anaphora. *Journal of Linguistics* 27: 301–35.

(1994) *The Syntax and Pragmatics of Anaphora: a Study with Special Reference to Chinese*. Cambridge University Press.

Kempson, R. (1986) Ambiguity and the semantics-pragmatics distinction. In C. Travis (ed.) *Meaning and Interpretation*, Oxford: Blackwell, 77–104.

Kiparsky, P. (1983) Word-formation and the lexicon. In *Proceedings of the 1982 Mid-America Linguistics Conference*, Lawrence: University of Kansas, Department of Linguistics, 47–78.

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

(1987a) Pragmatics and the grammar of anaphora: a partial pragmatic reduction of Binding and Control phenomena. *Journal of Linguistics* 23: 379–434.

(1987b) Implicature explicated? Commentary, *Behavioural and Brain Sciences*, 10 (4): 722–3.

(1989) A review of Relevance. *Journal of Linguistics* 25: 455–72.

(1991) Pragmatic reduction of the Binding Conditions revisited. *Journal of Linguistics* 27: 107–61.

(1992) Activity types and language. In P. Drew & J. Heritage (eds.) *Talk at Work*, Cambridge University Press, 66–100.

(1995) Interactional biases in human thinking. In E. Goody (ed.) *Social Intelligence and Interaction*, Cambridge University Press, 221–60.

(forthcoming) *Generalized Conversational Implicature*, Cambridge University Press.

Lyons, J. (1977) *Semantics*, 2 vols., Cambridge University Press.

McCawley, J. (1978) Conversational implicature and the lexicon. In P. Cole (ed.), *Syntax and Semantics*, vol. IX: *Pragmatics*, New York: Academic Press, 245–59.

Mehler, J., N. Sebastian, G. Altmann, E. Dupoux, A. Christophe & C. Pallier (1993) Understanding compressed sentences: the role of rhythm and meaning. In P. Tallal, A.M. Galaburda, R. Llinas & C. von Euler (eds.) *Temporal Information Processing in the Nervous System. Annals of the New York Academy of Sciences*, 682: 272–82.

Moravcsik, E. (1978) Reduplicative constructions. In J. Greenberg (ed.) *Universals of Human Language*, vol. III: *Word Structure*, Stanford University Press, 297–335.

Morgan, J. (1978) Two types of convention in indirect speech acts. In P. Cole (ed.) *Syntax and Semantics*, vol. IX: *Pragmatics*, New York: Academic Press, 261–80.

Papi, M. & J. Verschueren (1987) *The Pragmatic Perspective*, Amsterdam: Benjamins.

Reinhart, T. (1983) *Anaphora and Semantic Interpretation*, London: Croom Helm.

Searle, J. (1975) Indirect speech acts. In P. Cole & J. Morgan (eds.) *Syntax and Semantics*, vol. III: *Speech Acts*, New York: Academic Press, 59–82.

Sperber, D. & D. Wilson (1986) *Relevance*, Oxford: Blackwell.
(1987) Response to Peer review of Sperber & Wilson: *Relevance. Brain and Behavioural Sciences*, 10 (4): 697–754.
Visser, F.T. (1963) *An Historical Syntax of the English Language*, Part I, Leiden: Brill.
Wheeldon, L. & W.J.M. Levelt (forthcoming) Monitoring the time course of phonological encoding. *Journal of Memory and Language*.