5

# MINIMIZATION AND CONVERSATIONAL INFERENCE[1]

Stephen C. Levinson

*University of Cambridge*

## Minimization: A cross-cutting theme

The range of disciplines, methodologies and intellectual preoccupations represented at meetings like these must raise the question whether Pragmatics is not a mere flag of convenience under which divergent interests can momentarily find common profit in academic coalition. That suspicion is unlikely to be lulled by anything less than the odd successful synthesis across the different research traditions represented under the flag. It may be that in none of these traditions are the ideas yet clear enough to make successful integration possible; but I take it that we ought at least to be live to the possibility, and looking for opportunities to achieve such synthesis, and it is in this spirit that I offer the following tentative and in many ways premature suggestions in one small but important area.

In this paper I want to try and connect three distinct research traditions: *(Neo-)Gricean principles of inference*, *conversation analysis (CA)* and some recent issues in *syntax*. There is a particular interest in trying to connect the first two, for both Gricean approaches and CA approaches claim to be about the nature of *conversation*, and the lack of rapprochement ought to be embarrassing to our pragmatics coalition. In this paper, though, I do not intend to take on this synthesis of conversation analysis and conversational implicature directly. To do that, one would, for example, investigate naturally occurring speech events where the ground rules outlined in the maxims are not in force, and see whether, as predicted, implicatures fail to arise (see e.g. some suggestions in Levinson 1979).

Instead, I want to take a particular theme, *minimization in linguistic*

*expression and its inferential implications*, and to see whether we can find mutual support in those three fields (viz. the Gricean, conversation analytic and syntactic) for some general inferential principles. I have a subsidiary more specific theme, namely constraints on *reference to third persons*, but I have spread the net very much wider, necessarily I think, to gather in related issues, and to sketch the broader picture in which a resolution to that particular problem will hopefully fall out. That broader theme is the following apparent paradox: under certain specified conditions, *the less you say the more you mean*.

## 1. An anomaly in the Gricean programme and its solution: Informativeness in a Neo-Gricean framework

First, we must attend to a long standing anomaly in the Gricean programme that has only recently been recognized. In the heady early days of linguistic pragmatics, Gricean principles, and particularly the maxims of Quantity, were appealed to in order to solve numerous problems. Particularly important, for example, was the idea that the scope of negation was really a matter of pragmatic interpretation, with a pragmatic narrowing of 'external' to 'internal' negation; this made it plausible to consider presupposition essentially pragmatic (Allwood 1972; Kempson 1975; Atlas 1975). Meanwhile, Horn (1972) and then Gazdar (1976, 1979), were pointing out that there were highly systematic Quantity implicatures associated with certain scalar items like ‹all, some›, ‹know, believe›, where a stronger item entailed (in an appropriate sentence-frame) the weaker item, and the use of the weaker item implicated that the stronger item didn't hold. What was apparently not noticed for many years was that these two kinds of application of Gricean principles were in *fundamental conflict*: an inference from external negation to internal negation was a move from a semantically weaker to a semantically much stronger interpretation, while an inference from 'some' to 'not all', while an informational enrichment, was a highly restricted inference from the statement of one position to the *negation* of a stronger one.

### 1.1. *The clash between Q- and I-implicatures*

To make the clash of principles clear, consider the kind of reasoning thought to underlie the scalar implicatures like that from *some* to 'not all', as in:

(1) "Some of the miners voted for Thatcher"
(2) Not all of the miners voted for Thatcher

(3) Horn scale: <all, some>
(4) First Maxim of Quantity: "Make your contribution as informative as is required (for the current purposes of the exchange)"
(5) All of the miners voted for Thatcher

The reasoning goes roughly thus:

(6) The speaker in saying (1) has used the word *some* in contrast to the word *all*, with which it forms a 'Horn scale' (Horn 1972) of informativeness as in (3) (since 'all x have property P' $\vdash$ 'some x have property P'). Since the speaker has chosen a weaker, less informative expression where a stronger one of equal brevity is available, he would be in violation of the maxim in (4) if the stronger expression held. Therefore, he believes that the stronger statement (5) does not hold, and he has done nothing to stop me thinking that he so thinks, therefore he implicates (2).[2]

Now consider in contrast the inference from (7) to (8):

(7) "Bill turned the key and the engine started"
(8) Bill turned the key and then, as a direct and intended result, the engine started.
(9) Pseudo Horn scale: <since, and>
(10) "Since Bill turned the key (in order to start the engine), the engine started"

When we read (7) as meaning (8), we enrich what was said not only with notions of temporal sequence (as perhaps by Grice's maxim of Manner, "Be orderly"), but also with notions of causality, and also naturally with notions of intentionality or teleology. Now since there are causal connectives like *since* and *because*, and since these are semantically more informative than the truth functional connective *and*, we might, it may be thought, have a scale as in (9). If we now apply the reasoning we have just used for the scalar implicatures we will get *precisely the wrong prediction*:

(11) since the speaker has said (7) where there is the alternative *stronger* utterance (10) of more or less equal brevity (although the intentionality might need spelling out), he would be in breach of the maxim of Quantity in (4) if he was in the position to state the stronger utterance and failed to do so. Knowing this, he must implicate that (10) is not the case, and in particular that he does not know of any temporal, causal or teleological links between the two

clauses.

When Atlas and I (1981) investigated this clash we soon found that these were by no means isolated cases. There are of course large numbers of inferences that pattern like the <all, some> case; indeed all the scalar and clausal Quantity implicatures (like the inference from 'believe' to 'not know') investigated by Horn and Gazdar for example. Some of these are listed in (12) (recollect, of course, that these are only, in the manner of generalized conversational implicatures, default assumptions, which can be contextually cancelled). Intuitively partly what is going on here is a Saussurean value-by-opposition: where there are contrast sets of terms, the assertion of one term will be held to exclude the application of another term, even where the latter is logically compatible. Equally, though, there are large numbers of inferences that pattern the other way, like the enriched inference from *and* to 'since'. Some of these are listed in (14). Following Horn (1985), let us dub the classic Quantity scalar and clausal implicatures *'Q-implicatures'*, and (following Atlas & Levinson 1981) the counterposing inferences that are informationally richer the *'I-implicatures'*. (I use the symbol +> to stand for 'conversationally implicates'.)

(12) *Q-implicatures* (after Horn 1972, Gazdar 1979)

A. *Scalar*: given a Horn scale <S, W>, where the stronger expression substituted in an arbitrary sentence A entails the same sentence with the weaker expression, $A(S) \vdash A(W)$, and S & W are expressions of roughly equal brevity, then the use of W implicates the denial of the applicability of S: "A(W)" +> S knows ~A(S)[3]

Examples (omitting the epistemic qualification):

i. "Some of the boys were angels" +> not all were
ii. "Bill has written three books" +> not more than three
viii. "John is either a poet or a philosopher" +> not both
iv. "I often take sugar in my coffee" +> not always

B. *Clausal*: given a construction P which contains another clause q, where q is not entailed by P, and there is another construction R similar in meaning (and of roughly equal brevity) to P except that R does entail q, then the use of P(..q..) instead of R(..q..), implicates that the speaker is uncertain that q is the case:

Examples:

i. "I *believe* (cf. *know*) that John is away" +> He may or may not be

— speaker doesn't know which

ii. "I *thought* (cf. *realized*) it was a bomb" +> it may or may not have been a bomb — speaker doesn't know which

(13) *I-implicatures*

General form: given a pairing between a weaker expression W and a stronger one S in the same semantic domain, such that A(S) entails A(W), then if the speaker asserts "A(W)" he/she implicates the stronger statement 'A(S)' if that is compatible with what is taken for granted.

(14) *Examples of I-implicatures*

i. ('Conjunction buttressing'; Atlas & Levinson 1981)
"John turned the key and the engine started"
+> He turned the key *and then* the engine started (temporality)
+> He turned the key *and thereby caused* the engine to start (causality)
+> He turned the key *in order to make* the engine start (teleology)

ii. ('Conditional Perfection'; Geiss & Zwicky 1971)
"If you mow the lawn, I'll give you $5"
+> If and only if you mow the lawn will I give you $5

iii. ('Bridging': Clark & Haviland 1977)
"John unpacked the picnic. The beer was warm"
+> The beer was part of the picnic
"John was put in the cell. The window was barred"
+> The cell has a window

iv. ('Membership categorization devices'; Sacks 1972)
"The baby cried. The mummy picked it up"
+> 'The mummy' was the mother of the crying baby

v. ('Inference to Stereotype'; Atlas & Levinson 1981)
"John said 'Hello' to the secretary, and then he smiled"
+> John said 'Hello' to the female secretary, and then he (John) smiled

vi. ('Mirror maxim'; Harnish 1976: 359)
"Harry and Sue bought a piano"
+> They bought it together, not one each

vii. ('Demons', 'Frames', etc.; Charniak 1972, etc.)
"John pushed the cart to the checkout"
+> John pushed the cart full of groceries to the supermarket checkout in order to pay for them, etc.

viii. (Preferred coreference; see section 3. below)
"John came in and he sat down"
+> John came in and he, John, sat down.

Clearly, the I-implicatures are heterogeneous (ranging from referential specificity to 'bridging inferences', to 'frame' - based inferences), and much has been written on each kind.[4] The exact mode of inference presumably varies in each type, but the inferences do share some important properties: (a) they are more specific interpretations — what is implicated is a subcase of what is said; (b) unlike Q-inferences, this informational enrichment is not based on the negation of some stronger possible statement; (c) they are in potential conflict with Q-implicatures. Therefore we need an over-arching I-principle, which will *license* informative enrichments of these sorts; to produce such a principle, Atlas & Levinson (1981) set up the apparatus in (15) (in a somewhat simplified version):

(15) (a) *Maxims of Relativity*
"Don't bother to say what is non-controversial"
"Hear what is said as consistent with what is non-controversial"

(b) *convention of non-controversiality*
(i) it is non-controversial that referents and situations have stereotypical properties
(ii) the existence or actuality of what a sentence is 'about' is non-controversial ('aboutness' being characterized — Atlas & Levinson 1981: 42).

(c) *Principle of Informativeness*
The 'best' interpretation of an utterance is the most informative one consistent with what is non-controversial.

The effect of this apparatus is to induce a *more specific*[5] interpretation: *what is communicated is a sub-case of what is said*. Very often, the 'best interpretation' of what is said will be a more specific interpretation in line with *stereotypical expectations*.

It should be perfectly obvious that the Q-implicatures and the I-implicatures seem to operate on lines that look entirely inconsistent: the Q-implicatures induce the *negation* of the very sort of stronger interpretation that the I-implicatures actually appear to be promoting! This is because the first maxim of Quantity states, in effect, that "if the speaker didn't say an informationally richer utterance, he didn't mean it"; while the principle of Informativeness states, in effect, "go for the most specific interpretation in line with stereotypical expectations".

Horn (1985) in a recent article has further explored this clash between

inferential principles. He suggests that, Quality aside, Grice's maxims can be reduced to two, what he calls the Q-principle and the R-principle "following the current coy style that has given us D-structures and S-structures [...], I am using Q and R to *evoke* Quantity and Relation [Relevance] while leaving open the extent to which my principles map onto these two [Gricean] maxims" (1985: 13).

(16) (a) Q-Principle:
Hearer-oriented: "Make your contribution sufficient"
Or: "Say as much as your hearer needs, given R"
(cf. Zipf's 'Auditor's Economy')
(= Grice's first maxim of Quantity)

(b) R-Principle:
Speaker-oriented: "Say no more than your hearer needs, given Q"
(cf. Zipf's 'Speaker's Economy')
(= Grice's second maxim of Quantity, Relation, Manner)
(= Atlas & Levinson's Informativeness)

Here the Q-principle handles the scalar and clausal implicatures attributed to the first maxim of Quantity and illustrated in (12), and the R-principle handles the informational enrichment typical of the inferences in (14). I shall not follow this terminology because I do not think that Grice's second maxim of Quantity, our I-principle, can in fact be conflated with Relevance, but I shall discuss this below (see 1.5).

To further our understanding of the antinomic interaction between the Q- & I principles, it will help to tease apart the distinction between what each principle enjoins the *speaker* to do vs. what it licenses the *addressee* to think. I have done this in sketch form in (17) and (18):

(17) *Q-Principle*

1. *Speaker's maxim*: "Make your contribution as informative as is required for the current purposes of the exchange". Specifically: don't provide a statement that is informationally weaker than your knowledge of the world allows, unless providing a stronger statement would contravene the I-principle
2. *Recipient's corollary*:
Take it that the speaker made the strongest statement consistent with what he knows, and therefore that:
(a) if the speaker asserted A(W), and ‹S, W› form a Horn

scale, then one can infer K ~ (A(S)) i.e. 'the speaker knows that the stronger statement would be false'

(b) if the speaker asserted A(W) and A(W) fails to entail an embedded proposition q, which a stronger statement A(S) would entail, and S & W are 'about' the same semantic relations (form a contrast set), then one can infer: ~ Kq, i.e. Pq, P ~q (i.e. The speaker doesn't know that q obtains, or equivalently, it is epistemically possible that q or that not-q obtains)

Now let us try to do the parallel dissection for the I-principle. A first approximation is in (18):

(18) *The I-Principle*

1. *Speaker's Maxim: The maxim of Minimization*
   "Say as little as necessary" i e. produce the minimal linguistic clues sufficient to achieve your communicational ends, bearing Q in mind
2. *Recipient's corollary: Enrichment Rule*
   "Amplify the informational content of the speaker's utterance, by finding a *more specific* interpretation, up to what you judge to be the speaker's m intended point"[5]
   Specifically:
   (a) Assume that stereotypical relations obtain between referents or events, *unless* (i) this is inconsistent with what is taken for granted, or (ii) the speaker has broken the maxim of Minimization, by choosing a prolix expression
   (b) Assume the existence or actuality of what a sentence is 'about', if that is consistent with what is taken for granted
   (c) Assume referential parsimony — avoid interpretations that multiply entities in the domain of reference; specifically, prefer coreferential readings of reduced NPs (pronouns or zeros).

(I shall try to justify condition (c) in 3.2.1. below.) It is this I-principle that will be the focus of what follows, so let me draw the reader's attention to the obvious, but far-reaching, fact that a speaker's maxim of *minimization* ("say as little as necessary") has as immediate corollary an addressee's maxim of inferential *maximization* ("infer as much as necessary"). We can begin, I hope, to see some sense in our initial paradox: "the less you say the more you mean".

### 1.2. *Resolutions of the clash*

But what we must attend to here is how two such inconsistent bedfellows as the Q- and I-principles can manage to coexist. First, let us consider Atlas & Levinson's suggestions about how the two can be reconciled, outlined in (19).

(19) *The resolution of the clash: Atlas & Levinson's suggestion*

a. The Q-principle has limited domain of application, specifically, it operates only where there are clearly defined contrast sets, of which the Horn scale is prototypical. For <S, W> to constitute a Horn scale, where S is an expression informationally richer than W:
(i) S and W must be equally lexicalized,
(hence, there are no Horn-scales <iff, if>, <~, ~ >, (i.e. internal negation, external negation), — if there were, Q-implicatures would defeat 'conditional perfection' and presuppositional interpretations);
(ii) S and W must be 'about' the same relations, or be drawn from the same restricted semantic domain, (hence there is no Horn scale <since, and> to defeat 'conjunction buttressing')

b. Both the Q- and the I-principle are limited by the requirement of consistency with what is non-controversial or taken for granted.
(Hence there is no Q-implicature from "I cut a finger" to 'I cut someone else's finger'; and no I-implicature of conditional perfection from "If the door is locked I have a key in my pocket" to 'Iff the door is locked, I have a key in my pocket')

c. Where, nevertheless, there is a genuine clash, Q-implicatures defeat I-implicatures *unless* the Q-implicatures are inconsistent with what is taken for granted.

In effect, what Atlas & Levinson suggest is that a lot of the clashes observed above disappear on a proper analysis of what it takes to induce a Q-implicature. Further ones disappear when the general evaporation of implicatures in the face of taken-for-granted fact is taken into account. Finally, where a genuine clash does appear (as in c.), *Q wins over I*.

Atlas & Levinson's suggestions for resolution should be compared to Horn's (1985), outlined in (20), with applications in (21).

(20) *Horn's Principle of the Pragmatic Division of Labour:*
*I-inferences* induce stereotypical interpretations; therefore if a simple *unmarked* expression U denotes the set of extensions E, U will tend to become associated with stereotypical subset F of extensions E.
*Q-corollary*: if U I-implicates restriction to subset F, the use of an alternative, usually *marked* (more prolix or complex) term M will Q-implicate the complement of F, the non-stereotypical subset $F^-$ of extensions E.

(21) a. "secretary" I-implicates 'female secretary'
"emanuensis" Q-implicates 'male secretary'
b. "The Spaniards killed the Aztecs" I-implicates 'The Spaniards slaughtered the Aztecs'
"The Spaniards caused the Aztecs to die" Q-implicates 'The Spaniards indirectly (e.g. through disease) caused the death of the Aztecs'
c. "The bus comes often" I-implicates 'The bus comes often as far as stereotypical buses are concerned'
"The bus comes not infrequently" Q-implicates 'The bus comes less than often'
d. "Larry stopped the car" I-implicates 'Larry stopped the car in the normal manner with the foot brake'
"Larry caused the car to stop" Q-implicates 'Larry stopped the car in an unusual manner, e.g. by using the hand-brake'

We now have an explanation for the following sort of pattern:

(22) a. "John could solve the problem"
b. John solved the problem (I-implicature)
c. "John had the ability to solve the problem"
d. It's possible that John didn't solve the problem (Q-implicature)

In (22)a., given an assertion that A could do X, the I-principle induces the assumption (22)b.) that A did X, presumably on the basis that abilities are stereotypically manifested. In (22)c. the substitution of the marked and prolix paraphrase 'A had the ability to do X' induces a Q-implicature ((22)d.) that, as far as the speaker knows, it is possible that A didn't do X.

Horn's resolution schema basically assumes that the I-principle is generally prevalent, but that given a stabilized (generalized) I-implicature, Q then induces a contrastive implicature from the use of a term that is itself contras-

tive with the one associated with the stereotypical I-implicature.

Now there may be some doubt that the Q-principle that Horn is appealing to in his 'division of labour' is actually the same as the Q-principle involved in the derivation of the classic scalar and clausal Quantity implicatures. In particular, those classic inferences make no essential reference to the *manner* of expression, except to require contrasts between terms of roughly equal brevity; for they are based essentially on relative *informativeness*. But the Q-implicatures in Horn's 'division of labour' on the other hand do make essential reference to the brevity vs. prolixity of the manner of expression. They might thus be derived not from a Quantity maxim but rather directly from one of Grice's maxims of Manner, "Be brief; avoid prolixity".[7] I shall take up this matter below, but to distinguish these implicatures while equivocating as to their origin, let us dubb them (unforgiveably) the Q/M implicatures (thus the examples labelled 'Q-implicatures' in (21) and (22) we shall now call Q/M implicatures).

There are plenty of further puzzles, but the idea is, I trust, clear enough to allow us to proceed.

It is now evident how to put together the Atlas & Levinson resolution schema and Horn's alternative resolution to the clash between Q- and I-principles. Let us spell the *revised resolution schema* out as the three ordered conditions in (23):[8]

(23) *Revised resolution schema*
(i) genuine Q-implicatures from tight Horn scales and similar contrast sets of equally brief, equally lexicalized linguistic expressions 'about' the same semantic relations, take precedence over I-implicatures.
(ii) In all other cases, the I-principle induces stereotypical specific interpretations, *unless*:
(iii) there are two (or more) available expressions coextensive in meaning, one of which is unmarked in form and the other marked in form. In that case, the unmarked form carries the I-implicatures as per usual, but the marked form Q/M-implicates the non-applicability of the pertinent I-implicatures.

### 1.3. *Constraining the potential explosion of I-implicatures*

What this resolution schema does is use Q- and Q/M-implicatures to *constrain* the over-application of the I-principle. Nevertheless, even though all the I-principle does is induce *more specific* interpretations of statements that

are general over various possibilities, yet it doesn't take much imagination to see that the I-principle could induce an *inferential explosion*. From "John turned the key and the engine started" we might infer not only temporal, causal and teleological connections, but also stereotypical 'frame' like inferences about motor cars and their starting procedures; and if that, why not right down to the details of solenoids and cylinder-firing orders? While in certain circumstances it is quite probable that addressees are indeed m-intended to infer such specificities, certain constraints are clearly in order. I have some detailed suggestions but for now let me just say that it seems to work like this: unless conditions (i) and (iii) in the resolution-schema prohibit, I-implicatures can be as rich and specific as seem consistent with (a) an unreduced maxim of Relevance (more below), and (b) what, given the assumed state of mutual knowledge, the speaker might reasonably have m-intended.

### 1.4. *Minimization and concepts of informativeness*

We have played fast and loose with concepts of minimization, and concepts of informativeness. Let us turn first to examine notions of *minimization* that might be involved in some maxim like Horn's "Say no more than you must". Horn explicitly identifies his R-principle both with Atlas & Levinson's I-principle and with Zipf's (1949) least effort principle. But Atlas & Levinson's principle is based on a *semantic* notion: the best interpretation of an utterance is a proposition *more specific* than what was said; or, to phrase it as a maxim (which Atlas & Levinson refrained from doing), "speak in terms that are semantically general; don't bother to be specific". Let us call this *minimization 1* (see (24)). Whether being semantically general requires less effort than being semantically specific, and in what sense of 'effort', are questions that hardly arise here. In contrast a Zipfian least-effort principle seems to make most sense as some quantification over units of speech production, whether phonemes, morphemes, words or larger units. Let us call this *minimization2* (as in (24)). Such a quantification will be anything but straightforward, and will be complicated further by the fact that the *channels of communication* appear to be so ordered that displacement from the linguistic into the kinesic or paralinguistic will count as minimization (about which, more later).

(24) *minimization1*: semantically general expressions preferred to semantically specific ones;
*minimization2*: 'shorter' expressions (with less units of speech production) preferred to 'longer' expressions.

Horn gives no indication that he even recognizes the equivocation over these two rather different senses of reduction or minimization, and he certainly doesn't explicitly connect them. Moreover, a careful examination of his 'division of labour' between his Q- and R-principles will show, as already mentioned, that whereas the R- (our I-) principle is assumed to operate mostly in terms of semantic informativeness, the Q princple just here is operating on a measure of surface complexity or length; elsewhere, in Horn-scales, for example, it is quite clear that the Q-principle is operating primarily in terms of semantic informativeness.

The sceptic might at this point claim that there is a fatal equivocation here, between *'minimization1'* expressed in terms of semantic generality and informativeness and linked directly to Grice's second maxim of Quantity, and *'minimization2'* expressed in terms of units of speech production (and/or processing effort) which is properly linked to Grice's maxims of Manner (especially to the third maxim of Manner: "Be brief — avoid unnecessary prolixity"). However, there are at least three reasons to think that Horn's conflation is not totally unwarranted:

(1) There's obviously a general correlation, even if no necessary connection, between minimization of speech units and minimization of coded information. For example, a zero-anaphor is both informationally reduced (in that, unlike pronouns it can encode no person/number/gender information) and of course reduced from the point of view of speech production.

(2) There is a Zipfian argument to the effect that terms of wider application (i.e. semantically general terms) will tend to be used more often, and thus be shorter (q.v. Zipf's Principle of Economic Versatility ((the more semantically general, the more use)) and the Law of Abbreviation ((the more use, the shorter)) — Horn 1985: 30). The generally abbreviated form of pronouns (compared to the full NPs they stand for) is an exemplification.

(3) But above all, there are precise parallels in the inferential behaviour of forms minimal in sense 1 and minimal in sense 2. Thus the inferences to stereotypical specificity associated with the I-principle seem to be included both by brevity and by semantic generality, mostly of course together. Similarly, for the Q-principle: the classic scalar (and clausal) Quantity1 inferences are upper-bounding inferences: from the *absence* of an informationally stronger expression, one infers the *negation* of its applicability; in just the same way, the Q/M implicatures involved in the Hornian division of labour induce, from the *absence* of a briefer (unmarked) expression, the *negation* of the applicability of that shorter expression.

In what follows I remain ultimately agnostic over whether this subsumption of both minimization1 and minimization2 within the scope of an I-principle is legitimate. If it is not, then there is no reason to think that all the arguments below cannot be recast in a reasonably straightforward way in terms of the interaction of *three* independent principles, and their Q-, I- and M- (for Manner) implicatures.

We have said that the I-principle induces more informative interpretations of utterances. But what exactly does *more informative* mean? There have been many, but no fully satisfactory, suggestions. It will be expedient for what follows to take as a necessary condition for the notion of semantic informativeness the entailment analysis in (25); there are many well-known reasons why this is not sufficient. In addition Bar-Hillel and Carnap (1952) have operationalized a Popperian version (in which (25) holds) and have shown how to relativize the notion to contextual assumptions.

(25) *Entailment analysis of Informativeness*
A is *more informative* than B iff:
The set of entailments of B is properly contained in the set of entailments of A

It is the analysis in (25) that is usually assumed to be a *necessary* condition on the Horn scales and clausal alternates which are responsible for the classic Q-implicatures to the negation of the more informative statement. This definition will also do for now as a necessary condition on I-implicatures: if p I-implicates q, then q must entail p. However, in neither the Q- nor the I-induced implicatures is it anything like a *sufficient* condition: in both cases there are strict constraints requiring a semantic relatedness closer than that captured by mere meaning-inclusion. In the case of the Q-implicatures, inferences arise from items that form a *contrast set* over a limited semantic domain, while in the case of the I-implicatures the implicature that entails the sentence uttered must be *more specific* than what was said. In trying to spell these notions out, we rapidly get into deep theoretical water; here we must pass on, but the issues clearly require further attention.

### 1.5. *Rethinking the maxims*

What implications do these proposals have for our overall view of the maxims of conversations? First, there is the obvious question whether the I-principle is a *co-operative* maxim at all: perhaps, as Horn (1985) in places implies, it is merely the instantiation, in the realm of linguistic inference, of the Zipfian universal principle of *Least Human Effort*.

Kasher (1977) in fact has suggested that all the maxims follow from a principle of speaker self-interest: "Given a basic desired purpose, the ideal speaker chooses that linguistic action which, he believes, *most effectively and at least cost* attains that purpose" (cited in Sampson 1982: 205).

I do not think, though, that the reduction to self-interest is quite that easy, even in the case of the I-principle. First, there is plenty of evidence that the problem might be the other way around: that people, under many circumstances, like to speak at length – what else would motivate the elaborate turn-taking machinery of conversation, that seems specifically devised to constrain a natural tendency to 'hog the floor' (see 2.)? From that perspective, of *pressure on the 'floor'*, a co-operative maxim of minimization does not look so redundant at all. Secondly, the I-principle isn't really just a simple instantiation of the Zipfian principle of Least Effort, for it operates not just on some quantification of articulatory effort, but also and primarily on the *semantic or information level*, preferring the general to the specific, even where no articulatory reduction will thereby be achieved, witness (26):

(26) a. Ann came in. *Ann* sat down
b. Ann came in. *She* sat down

If b. is preferred to a. (on a coreferential interpretation) by the I-principle (since the pronoun is more semantically general, or at least has wider application, than the proper name), this is not going to be accounted for either in terms of reduction of articulatory effort or in terms of a reduction of cognitive effort (which would surely be increased by having to find an alternative specification for the same referent).

I shall therefore continue to assume that any such principle as the I-principle would indeed need to be independently motivated, and presumably so too for the Q-principle, and that Grice's co-operative source is not obviously wrong. In addition, whatever the source, it must be relatively constant and normative in order to account for the generalized nature of the implicatures under discussion.

A second important question that arises is whether, given the Q- and I-principles, we need any further maxims at all. Horn (1985) specifically favours such *reductionism*, and Wilson & Sperber (1981) take it one stage further and reduce all to one, their maxim of Relevance. Note that neither reduction seriously attempts to reduce Quality; the focus is on the reduction of the rest of the maxims. What Horn goes on to do is collapse the second maxim of Quantity, our I-principle, with Relevance, to create his R-princi-

ple, assuming that the Manner maxims will fall either under the first maxim of Quantity (his Q-principle), as in the case of 'Avoid ambiguity', or under his R-principle, as in the case of 'Avoid prolixity'.

There are reasons I think to doubt that one can dispense with the Manner maxims governing the *superficial form* of what is said by subsuming them within maxims that govern *informational content*. But I leave that aside here. For the interesting reduction is the conflation of the second maxim of Quantity ('do not say more than is required') and Relevance ('be Relevant'). The argument is: "What would make a contribution more informative than required, except the inclusion of material not strictly relevant to and needed for the matter at hand?" (Horn 1985: 12). Exactly similar sentiments have been voiced by Wilson & Sperber (1981). It is therefore important to consider in detail whether the constraints on Relevance are indeed of an informational character.

But before considering this, let us turn to the other reductionist attempt by Wilson & Sperber (1981), who try to conflate all four maxims into the one developed notion of Relevance (I shall call this SW R to avoid ambiguity).[9] SW R as a maxim is the injunction to make one's utterance maximally SW R, where SW R is a vector value of most contextual implications balanced by least processing effort. Their reduction of Quality just seems to me to be spurious; in any case it is not central to their claims. Their reduction of the maxims of Manner are at least partially promising, and here they are in the same camp with Horn, and the present proposals, in assuming that 'Be brief' and 'Be orderly' reduce at least partly to maxims governing informational content. (As I've hinted, though, I think it will be more difficult to do away entirely with constraints over surface form in favour of constraints over information content, but since SW R contains a measure of processing effort, this may perhaps allow more scope for reduction here than our own rival proposals.)

The maxims of Quantity reduce more directly to SW R, because the latter contains a measure of informativeness; indeed in effect it *is* a maxim of Quantity, purportedly balancing maximum information (Grice's Quantity1, our Q-principle) against processing effort (which will partially achieve the constraints in Grice's Quantity2, our I-principle).

In actual fact, though, it is most doubtful that a single principle like SW R can hope to cover the range of both the I- and Q-inferences, their clash and their resolution. Consider SW R as a maxim: 'The speaker tries to express the proposition which is the most relevant one possible to the hearer' (Sperber &

Wilson 1982: 75), where 'most relevant' is to be construed as maximizing the contextual implications while minimizing the cognitive effort involved in drawing them. This maxim ought to force the speaker to be explicit, if *express the proposition* means 'say' in Grice's sense. Therefore, it should work exactly against the direction of our I-principle: speakers should always speak just as specifically as they mean to be interpreted, presumably inducing Q-type inferences from the failure to make a stronger statement. If on the other hand *express the proposition* means 'say and/or implicate', then SW R would have similar effects to our I-principle (although SW R inferences would be much less constrained). It is only on this interpretation that SW R might hope to capture the classic Relevance inferences, as for example in partial answers to questions. I expect that there is a deep equivocation in the SW proposals here, and it is only this equivocation that makes it possible for their SW R to apparently cover both our I- and Q-principles, but I cannot develop this in detail here.[10]

Let us now return to the crucial issue whether maxims that are basically measures of informativeness can subsume a maxim of Relation or Relevance. First, there are some real reasons to think that the core intuitions of 'connectedness' that seem to underlie the ordinary use of the term *relevance* are not necessarily of an informational character (see Holdcroft in this volume, for independent remarks along the same lines). Consider the examples in (27):

(27) a. A: Hello
B: (i) Hello
(ii) The etymology of that word is now believed to be from the French 'hola'

b. ((In ticket office)) (after Allen 1983)
A: When's the train to Windsor leave?
B: (i) 10.15 on platform 10
(ii) 10.15

c. A: ((after identifications on the telephone))
Where can I reach John?
B: ((operator in building))
(i) I'll put you through
(ii) On extension 1423

d. A: Do you stock LT 188?
B: (i) Big or small?
(ii) Yes

In each of these, there is a clear intuition, I hope, that the first response labelled (i) by B is *more relevant* than the second labelled (ii), even though, except in case b., the second responses display more connectedness of information. In b., by Horn's argument ('saying more than necessary will be irrelevant'), one might expect the (ii) response to be more relevant (the platform specification not being asked for), whereas in fact (i) is (see Allen 1983).

What seems to be going on here is that the ordinary notion of 'relevance' is constructed out of at least two notions of connectedness that are not essentially informational:

(i) *rule-specified adjacency* of two kinds of speech acts or activities, e.g. greetings, or adjacency-pairs in general (see e.g. Sacks 1972, on 'relevance rules'); thus what is 'relevant' is what should be done next;

(ii) *teleological connectedness*: a response is 'relevant' to the extent that it meets the other's needs or interactional goals (as in c.-d.); thus what is 'relevant' is what is wanted or required. Conversation analysis (CA) has explored in detail *sequential expectations* and the inferences that these expectations give rise to, and it is in this area of study, it seems to me, that a proper account of these primary aspects of relevance will be found.

Yet a third notion that seems to be involved in the pretheoretical notion of relevance is *topical connectness*. An important 'lay' notion, it has proved very resistant to theoretical understanding. But from CA work, it seems likely that notions of 'topic' in conversation are not fundamentally linked to notions of informativeness; loosely, we can say that 'topic' is not related to *what* is being talked about, but rather in what *light* it is being talked about (a concept explicated in Schegloff 1972 in terms of 'formulation').

None of this is to deny that there are very close links between notions of informativeness and notions of relevance; there is indeed *one* kind of connectedness which is a connectedness of information, or 'aboutness'. But there are certainly other kinds, like the connectedness of recurring conversational sequences and the connectedness of purposes and goals, which are certainly unreduced by any of the proposals reviewed.

To sum up: if we assess the recent proposals to reduce the maxims in the light of the problems raised by the clash of what I have called Q- and I-implicatures, we are more than likely to be inclined towards the judiciousness of Grice's (1967, 1975) original set of maxims. His Quality seems unreducible, whether or not it is properly part of this scheme; his Quantity1 is our Q-principle; his Quantity2 is our I-principle; his maxim of Relation looks at least partially unreduced; and there are at least indications that we may need sep-

arate maxims of Manner.

### 1.6. *Summary*

We started by noting an anomaly in the Gricean programme: the 'radical pragmatics' attempts to reduce presupposition to implicature in fact rested on a species of inference quite distinct from the kind they had been attributed to (namely those due to the first maxim of Quantity). When explored, this species (I-implicatures) turned out to have numerous representatives. Moreover, on the face of it, these highly informative inferences are in direct conflict with those properly attributed to the first maxim of Quantity. An anomaly of these dimensions might have signalled the early demise of the Gricean programme. However, the clash turns out to be less direct than it at first seems, and there appear to be clear resolution rules organizing the interaction of the two antinomic principles. Building on ideas in Atlas & Levinson (1981) and Horn (1985), we were able to propose a detailed resolution schema for potentially conflicting Q-, I- and Q/M-implicatures.

We have spent considerable time investigating the nature of the underexplored I-principle, and in what follows it will be to the fore. In particular, of central interest to us will be its somewhat paradoxical implication that *minimized forms get maximized interpretations*. Or to put it in slogan form, the less you say, the more you mean!

## 2. Principles of minimization in conversation analysis

In section 1., we have developed some hypothetical, abstract 'conversational principles' in order to account for certain observable generalized conversational implicatures. But any theorist positing such principles is open to the charge of ad-hocery unless the same principles can be empirically shown to operate in actual conversation. In short, a conversational principle had better show up in conversation! Otherwise, given any set of puzzling pragmatic inferences, we can just invent an ad hoc maxim to account for them; instead, we should insist, of any Neo-Gricean theory, that the principles hypothesized can be shown to guide actual conversational activity. This is not easy, but in this section I attempt to show that the principles in section 1. are in fact well-grounded in the details of conversational organization. To do so, I draw widely on work in conversation analysis (CA) that attends in a general way to the handling of information in conversation, to see if we can discern any analogues of our Q- and, especially, our I-principles.

### 2.1. *Information and informativeness in CA: A social economy of information*

Observers of work in conversation analysis have been especially impressed by detailed findings about the mechanics of conversational practice — turn-taking and repair, for example. It is less obvious that CA has been fundamentally concerned with inference since its inception, although a glance at the work of Garfinkel or Sacks will confirm this; and much recent CA work is directly concerned with principles of inference, with explicating how what is not said can be implied But why not say it?

Here work in CA has shown that information is a valued and guarded resource (observations that fit with much in the ethnography of speaking, e.g. Keenan 1976), indeed that there is a veritable *economy* of information, with social constraints operating in the categorization of information as 'news' vs. mere 'information', 'news' as 'good' vs. 'bad', and procedures governing the disclosure of each kind of category to different categories of others (see Sacks, 19 Oct 1971). Recollect, for example, Sack's (1975) demonstration that 'everyone has to lie'; this follows from the fact that answers to the conventional opening question "How are you?" are constrained in that (a) an honest answer might lead away from the business in hand (and the kind appropriate to the interlocutor), and (b) news about troubles and successes ought to go in a well-ordered sequence, to one's nearest and dearest first. On the other hand, failure to tell some people some things can be seen as a serious delict, this motivating the particular structure of closing sequences of telephone conversations which allow matters other than the main topic of the call to be inserted in the final stages. Such matters may be pressing in the sense that they have to be told, and have to be told in the absence of parties who are not appropriate recipients of this telling. Thus in the following telephone call, after making arrangements to meet, which is typically the immediate precursor to the closing of the call, one such crucial piece of news is inserted:

(28) (Schegloff & Sacks 1973 (1984: 93); transcription simplified)
B: Fine.
A. *Other* than that don't uh
[
B: Fine
A: Don't bother with anything else.
→ I huh::: (1.2) I-uh::: I *did* wanna tell you, en I didn't wanna tell

you uh:: last night. Uh because you had entert- uh, company. I-I-I had something- *terrible* to tell you. So uhh

[ [

B: How terrible *is* it? How terrible *is* it?

A: Uh, tuh- as worse as it could *be*.

(0.8)

B: W-y' mean Ada?

A: Uh yah

B: Whad' she do, die?

B: Mmhmmm

Such examples point up the limitations of some simple minded application of the first maxim of Quantity, 'say as much as is required', for the transmission of information is of course socially and interactionally constrained.

### 2.2. *The interactional distinction: Informer/non-knower*

Clearly, such an economy of information, with rules prescribing and proscribing transmissions, presupposes a division, for each unit of information, between *knowers* and *non-knowers*. Such a division, Heritage (1984b: 309-10) reminds us, is presumed by the asker of a question, and the slot after an answer can be expected to indicate whether the questioner's state of information has been changed, the typical indication being the particle *Oh*:

(29) (Heritage 1984b: 308; simplified)

A: Well listen (.) did you didju phone your vicar yet?

(0.3)

B: No I ain't

(0.4)

→ A: Oh:

The frequency of this *Oh* after answers contrasts with its absence in classroom question-answer sequences (where non-informing exam questions are the rule) and in courtroom interrogation by a neutral party where *Oh* might be taken to imply that the neutral party acknowledged the acquisition of a fact (Atkinson 1979, cited in Heritage 1984b: 339).

Heritage (1984b) goes on to show that there are a set of alternates in the slot after an informing; plain *Oh*, for example, after announcements contrasts with *Yeah* which is only the weakest acknowledgement of an informing, and often presages statements of prior knowledge:

(30) (Heritage 1984b: 305)
H: Listen, Bud's alright
→ J: Yeah, I know, I just talked to 'm.

Here we have something like a Q-inference: although *Yeah* might be thought to be compatible with the receipt of new information, by a contrast with the stronger receipt token *Oh, Yeah* implicates failure to inform.

But the main point here is that use of such particles (as also the use of questions) presumes a distinction between the key roles of informer and non-knower, roles that somehow have to be attributed in interaction, even if, as the substitution of *yeah* for *Oh* in (30) makes clear, these sometimes need to be interactionally adjusted. Just how people can predict so accurately what others do and do not know is yet another mystery of interactional dynamics.

### 2.3. *A maxim: 'Don't tell people what they already know'*

Being able to distinguish between knowers and non knowers of a piece of information is a precondition to being able to obey the 'maxim' (Sacks' term) "*Don't tell people what they already know*", a special subcase of which is the rule "Don't tell people what you've already told them" (Sacks 1971, Oct. 19: 9).

One strong kind of evidence for such a maxim can be found in the *pre-announcement sequence* (Terasaki 1976), where a preliminary turn checks out whether the recipient has or has not heard some news. Stereotypical versions are the joke pre-announcements – "Have you heard the one about the Pink Martian?"; more businesslike is example (31):

(31) (Terasaki 1976: 28)
D: Didju hear the terrible news?
R: No. What.
D: Y'know your grandpa Bill's brother Dan ...

One general motivation, although not the only one, for such preannouncements, is that by prefiguring what is coming up, and providing for the recipient to indicate that the news is already known, they allow a telling to be aborted. (How they effectively prefigure what is coming up next is a fascinating area of its own; see Terasaki 1976; Levinson 1983: 349ff.)

Another kind of evidence of sensitivity to recipients' knowledge states comes from multi-party conversation. Where there are more than two participants, the possibility arises that information may be differentially distributed between two or more recipients. Thus in the extract in (32), A produces a pre-

announcement, B the expected 'go-ahead', but C simultaneously disaffiliates from B's response indicating that C already has possession of the 'news', and this is sufficient to hold up the telling:

(32) (Heritage 1984b: 304)
A: Hey we got good news
B: What's the good news?
[
C: I kno:w
(.)
A: Oh ya do::?
[
D: Ya heard it?

What is of interest is the apparent obligation to indicate, if one is already in the 'know', that one is not a possible recipient of this piece of news. An alternative would have been for C to say something like: "Yeah, go ahead and tell him", or to actively collaborate in the telling, all alternatives discussed in the literature (see e.g. Sacks 1971: Oct. 19th). Once again, then, evidence for the rule 'Don't tell people what they already know'.

Again, in multi-party conversation, one can sometimes find that a speaker, having found an intended recipient to be non-attentive, will turn to find another, and in so doing must recast what he says to avoid telling the new recipient what would have been news to the prior one. Goodwin (1979) subjects the segment in (33) to this kind of scrutiny, using the speaker's gaze direction as indication of intended recipient. (The direction of gaze is marked thus: A → B indicates that A looks at B; the duration is marked as a line under the concurrent speech.)

(33) (Goodwin 1979: 98-99; considerably simplified)
J: I gave, I gave up smoking cigarettes::
J → D ----------
D: Yea:h
(0.4)
J: I-uh one-one week ago t'da:y acsually
J → B ----------
J → A ----------

Goodwin argues that J begins by selecting D as recipient for his news, but finds D pretty uninterested (*Yeah*, in contrast to *Oh*, hardly being the kind of congratulatory token one might expect after such an announcement). J there-

fore turns to find another recipient, selecting (and thus gazing at) B. But B is J's spouse, and there's every reason to supppose that B knows that J has given up smoking. So the transition from an unknowing recipient (D) to a knowing recipient (B) requires a recasting of the news — no longer as a mere fact, but as an announcement of an 'anniversary', the end of the first week of non-smoking. Such a recasting is entirely appropriate: someone who already knew John had given up smoking would be the very person best able to appreciate that he has now passed the seven day landmark. (This interactional adjustment of information within the course of a single turn in response to interactional feedback or the lack of it is a recurrent finding in recent CA work; cf. e.g. Davidson 1984.)

All these conversational procedures point to a pervasive orientation which "involves avoiding telling recipients what they already know" (Heritage 1984b: 303, 338 fn.8). That in turn has the really rather astounding implication that we try to keep for each significant possible interlocutor a kind of running tally of exactly what we've told each of them and when. Without that, a full-fledged economy of information would, one supposes, hardly be possible.

### 2.4. *Minimal specifications can get maximal interpretations*

Let us now turn to see whether we can find conversational evidence for the pattern we have attributed to the I-principle, namely the inferential richness associated with minimal (short) or semantically general linguistic expressions.

There appears to be a general practice, with its associated interpretive corollary, to omit specification where that would spell out what can be presumed by virtue of what is being described. A first example comes from Schegloff's (1972) analysis of formulations of location. He notes that in calls to a police department, where a location for investigation is given in terms of, say, the intersection of two streets, there will be a presumption that both parties are in the same city (the presumption leading occasionally to misunderstandings that can only arise thereby; 1972: 101). Thus, *zero-specification* of place at a hierarchical level *n*, seems to I-implicate (in my terms), the assumption of copresence in *n*. Similarly, Schegloff notes (1972: 102), if A and B are copresent in New York, a description that "John is in the East" (where *the East* (of the USA) would normally be taken to include New York) would implicate that John was not in New York. The reasoning seems entirely in line with Horn's 'division of labour': zero specification I-implicates co-presence; a lin-

guistic specification in marked contrast to zero but compatible with co-presence will Q/M-implicate the complement of the I-implicature — i.e. that John is not in New York.

In commenting on location specifications, Schegloff also notes that place terms can be richly suggestive of further information, for example place terms can be used to answer questions about what someone is doing ("he's at school"). Similarly "He's gone to the supermarket" suggests the usual one, the one we go to. These are of course stereotypical I-implicatures. It is of interest that, just as predicted by the I-principle, *reduction* of these terms reinforces the I-implicature: "John went to O school", "John's at O church" (with the un-English absence of the article), force an interpretation in terms of doing the stereotypical actions at those locations.

Another notable early attempt in CA to grapple with the informative reading of minimal descriptions is the paper by Sacks (1972) that addresses the two-sentence children's story in (34)a.:

(34) a. "The baby cried. The mommy picked it up"
b. Inferences:
i. The baby and the mommy are co-family members
ii. it is stereotypical that babies cry;
It is non-reprehensible that this baby cried
iii. It is stereotypical that mothers comfort babies;
the mother did the proper thing
iv. The event described in the first clause (E1) comes before the event described in the second (E2)
v. E2 occurred *because* of E1

To explain the associated inferences in (34)b., Sacks develops a quite elaborate apparatus. What in effect the apparatus does is spell out some of the operations involved in deriving a stereotypical I-implicature: the notion of a 'membership categorization device' (i.e. a set of lexemes like *baby*, *mummy*, *daddy* that label elements of a superordinate unit like the *family*) identifies referents as elements of some relevant set. Where that super-ordinate set is recurrently instantiated, like 'the family', then the strong interpretation is induced that both elements are from the same instance of the recurring superordinate set (as in inference (34)b.(i)). Where any stereotypical activities are mentioned one should hear them in accord with the stereotypes; and where two events are mentioned that could be linked by a social norm (e.g. 'mothers should pick up distressed babies to comfort them'), then one

should interpret those events as co-occuring *because* of the norm. This paper is, in fact, one of the most detailed attempts, together with some of the 'frames' approaches in A.I. (which it historically antedates), to spell out exactly how stereotypical interpretations are reached.

Another aspect of conversation that ties in closely with the I-principle is *preference organization*. The notion of *preference* as it organizes the possible responses to first parts of adjacency pairs fits, at least to some considerable extent, with the Hornian 'division of labour'. Pomerantz (1975, 1978, 1984) and others have noted that possible responses to certain prior utterances are clearly distinguished as being preferred or relatively dispreferred. For example, in response to invitations, acceptances are undelayed and brief, rejections hesistant and buttressed with appreciations and excuses; the following contrastive examples (from Atkinson & Drew 1979: 58) are canonical:

(35) A: Why don't you come up and see me sometimes
[
B: I would like to

(36) A: Uh if you'd care to come and visit a little while this morning I'll give you a cup of *co*ffee
B: Hehh Well that's awfully sweet of you, I don't think I can make it this morning. hh uhm I'm running an ad in the paper and-and uh I have to stay near the phone.

Dispreferred responses typically exhibit at least some of the following properties (see Pomerantz 1984, Levinson 1983: 332ff):

1. Delay — by pause, preface or displacement to other turn
2. Prefaces — particles like 'Well', 'Uh'; token weak agreements prefatory to disagreements; apologies or appreciations; hesitation; qualifiers of assertion
3. Accounts — reasons for dispreferred action having to be done
4. Only finally, the declining element.

This contrasts with the simple format typical of preferred responses.

As I've pointed out elsewhere (Levinson 1983: 333), this contrast between simple form for unmarked meaning, complex form for marked meaning, is entirely in line with the linguistic notion of *markedness*. And of course, it is also entirely in line with Horn's *division of labour* between the I- and Q/M-principles, which specifies that by I the unmarked form should implicate the informative, stereotypical response, while the marked form will Q/M-implicate the complement of the I-implicature. There are perhaps one or two doubts about the application of the Hornian resolution here which are

worth reviewing:

(a) Is the unmarked (preferred) form really the stereotypical and expected response? I think one can say that it is: whatever an individual recipient's feelings about, say, the receipt of an invitation, the preference format itself is so designed to make acceptance easier than refusal (one has, after all, to dream up apologies and excuses).

(b) Is there really an I-implicature involved, or are preferred responses always direct and fully explicit? Here consider, for example, a response to an invitation that goes "How lovely". Such an appreciation *without* other material must be read as an acceptance.

(c) By the I-principle and its maxim of minimization, shouldn't a pause be interpreted as doing the stereotypical preferred response, rather than as constituting the 'delay' component of a forthcoming dispreferred (as empirically it does)? Well, in the case of preferred responses, given that the minimal linguistic forms already carry the stereotypical preferred interpretations, the absence of *these* is quite clearly a *with-holding*. Thus, the absence of the form responsible for the I-implicature will, just as usual, Q/M-implicate the complementary interpretation. The only way in which such silences fail to fit the Hornian division of labour is that in this case the marked form is the most minimal form of all! But overall, preference organization would seem an impressive vindication of Horn's proposal.

We may ask why there is this favouring of maximal interpretations of minimal specifications in conversation? One answer may be a *general tendency to compression* that is promoted by the *turn-taking system* (Sacks, Schegloff & Jefferson 1978 [1974]; Schegloff 1981). For it seems that the basis of the turn-taking system in conversation is the 'turn-constructional unit', the minimal syntactic and prosodic unit upon whose completion the transition of speakership becomes immediately relevant. Since, unless the turn-taking system is suspended by joint agreement, one can only be certain of *one* such unit before loss of speaking rights, there is a motivation for compression of information. Again, in a similar way, since many *sequences are closed by acquiesence or agreement*, and since that type of response is generally preferred in the technical sense, there is a motive for compression — in a sequence that goes:

A: [Assessment]
B: [Agreement]

neither A nor B may get more than one such turn-constructional unit before

the matter is effectively completed.[11] But issues of sequence truncation bring us to our next topic.

### 2.5. *The implicit preferred to the explicit*

There are a number of well studied conversational practices that seem to favour *the implicit over the explicit* and thus seem entirely in line with a maxim of minimization. Good examples are provided by the interaction of preference organization and specific sequences. Where there is a clear sequence type, there is also usually a *canonical extended form* which may often in practice be truncated. Given a full canonical form and a series of more and more truncated forms of the same sequential type, then often there appears to be a preference for the truncated sequences over the full extended sequences.

Well-known in this regard is the tendency towards the truncation of repair-sequences, where the full form of other-initiated repair, forming a sequence of three turns, is dispreferred in favour of self-induced self-repair, thus avoiding the interactional negotiation in a sequence altogether. Where a failure to self-repair occurs, but a recipient can guess what was intended, then 'embedded' or covert repair seems to be preferred, in which a recipient waits until his turn and replaces the offending term with the 'correct' one on the first relevant occasion of reference. These patterns are too well-documented and too well known to require exemplification (see Schegloff, Jefferson & Sacks 1977).

A second kind of sequence truncation, detailed in Levinson (1983: 360ff), occurs with request sequences. Again the full extended normal form in (37)a. seems to be least preferred, in contrast to the truncated sequences in b. and c. (the last truncation providing the 'indirect speech act' interpretation):

(37) *Truncation of pre-request sequences*

(a) Position 1: Pre-Request "Do you have size C batteries?"
Position 2: Go ahead "Yes sir"
Position 3: Request "I'll have 4 please"
Position 4: Compliance ((turns to get))

(b) Position 1: Pre-request "(Are) you going to the party?"
Position 2': offer "Yes would you like a lift?"
Position 3': acceptance "Oh I'd love one"

(c) Position 1: Pre-request built to truncate ('Indirect Speech Act')
"You don't have this number I don't suppose?"
Position 4: Compliance "Sure. Cambridge 60385"

Another case of preference ranking over extended and truncated forms of the same sequence type occurs in self-identifications on the telephone (Schegloff 1979; I draw the examples from that source).

(38) (a) *Greetings plus self-identifications*

1 C: ((rings))
2 R: Hello,
3 C: Hi. Susan?
4 R: Ye:s,
5 C: This's Judith (.) Rossman
6 R: *Jud*ith!

(b) *Greetings with submerged self-identifications*

1 C: ((rings))
2 R: Hello
3 C: Hello
4 R: Hi

That the second sequence (b) is preferred to (a) is shown by:
(1) the fact that in turn 3 the caller C often produces a "Hello" without further self-identificatory material, and then, only after a pause provides a self-identification, as in (c):

(c) *Delay before self-identification*

1 C: ((rings))
2 R: Hello
3 C: Hello Charles.
(0.2)
This is Yolk.

In other words, there is evidence that the truncated sequence of type (b) is planned for, while the full extended form is resorted to only when the shorter form fails.
(2) The fact that there are intermediate forms between sequence type (a) and (b), particularly in the progressive upgrading of the third turn as in (d):

(d) *Third turn progressive upgrading*

3 C: Hello (.) It's me (.) Steve (.) Levinson

(3) The occurrence, sometimes, of a 'try-marked' other-identification in third turn:

(e) *Try-marked other-identification*

1 C: ((rings))

2 R: Hello
3 C: Hello *Ilse?*

Where this occurs with a low-rise intonation it apparently signifies not uncertainty about the identity of the recipient R, but uncertainty about whether R can recognize C. This form would be without motivation if it were not inviting R to make a guess at C's identity, the 'questioning' format providing R with the next turn to do that, while prefiguring a self-identification if R should fail. (4) Even if R fails to recognize C from the third turn containing a try-marked other-identification, self-identification may still be withheld, a further example of voice-quality alone being offered as in (f):

(f) *Self identification withheld in fifth turn*
1 C: ((rings))
2 R: H'llo?
3 C: Harriet?
4 R: Yeah?
5 → C: Hi!
6 R: *Hi:!*

In short, there is ample evidence, carefully documented in Schegloff (1979), that, if possible, self-identification should be done covertly, through the provision of minimal voice-quality samples. There is also evidence that this requirement can be progressively relaxed until identification is achieved.

A final speculation is relevant to self-identification sequences. We have already suggested (in 1.4 above) that communicative channels are ranked as far as minimization goes, so that the gestural channel counts as more minimal that the spoken channel (see also the comments on the Guugu Yimidhirr segments in 4.0 below). It seems likely too that on such a ranking the encoding of information in prosodics and paralanguage counts as more minimal than encoding in the linguistic channel proper. Since covert self-identifications are done by 'signatured' voice-quality displays (see Schegloff 1979 for details) and are preferred to lexical self-identifications, they fit this pattern.

**Summary so far**

Let me summarize so far. CA findings hint at an 'economy' of information replete with obligations and prohibitions on the transmission of information, and especially perhaps on the *mode* of transmission. But I've been keen to establish, in my review of some of this work, the operation of principles closely allied to our I-principle and the maxim of minimization — "Say little

and infer much". First, we've shown that there's interactional sensitivity to the attribution of the complementary roles of *informer* and *informee*. This in turn is the precondition to the operation of a rule "Don't tell people what they already know", such a rule informing many aspects of conversational organization. That rule is a subcase of our I-principle, Grice's "Don't say more than is required". That maxim of minimization can be seen to operate in two kinds of conversational organizations: minimal specifications supplemented with special inference rules, and the tendency towards the truncation of sequences.

### 2.6. *Evidence for the clash between I vs. Q: Reference to persons*

We have earlier outlined (in section 1.1) a speaker's maxim of Minimization, and the hearer's corollory, the Enrichment Rule; and we have shown how these combined into an I-principle run directly into conflict with the Q-principle.

I now want to suggest that these theoretical constructs are, happily, directly supported by some CA findings about *reference to persons*. The locus classicus is a condensed paper by Sacks & Schegloff (1979) where they claim to see two wide-ranging principles of conversational organization interacting in an interesting way. The principles are (a) a principle of *minimization* (their term), which in this reference domain amounts to a preference for monolexemic reference forms (or at least single reference forms), and (b) a principle of *recipient design*, which in this domain amounts to a preference, where possible, for a reference form that will allow the recipient to recognize that a mutually known referent is being referred to (a *recognitional* in the authors' parlance). Most of the time these two principles co-exist quite happily, being typically instantiated in English by the use of a *name* which is both a minimal form and a recognitional: Bill, Sue, Smith, etc. Note though that it takes the two principles to explain the preponderance of reference to persons by name: recognition alone might be achieved by elaborate description, but this would run counter to minimization; minimization might be achieved by indefinites like "someone" but this would run counter to recognition.

But the main evidence for the two principles comes from two kinds of occasions where they are in conflict.[12] In the first kind, the speaker is uncertain that the recipient will be able to recognize the referent, and so uses a special form of the minimal recognitional, marked by a rising intonation with a following pause which is provided for acknowledgement of recognition (the authors call these features a *try-marker*):

(39) (Sacks & Schegloff 1979: 19)
A: ... well I was the only one other than that the uhm tch Fords?, uh Mrs Holmes Ford?
You know uh the the cellist
[
B: Oh yes. She's she's the cellist
A: Yes

Evidently, what happens is that if there is no assent to recognition following a try-marked minimal form, there is a step by step escalation, from minimal name to augmented title plus names, and then from names to descriptions, until recognition is achieved. Minimization is thus successively relaxed in favour of Recipient Design & recognition, but that it still operates is shown by the reluctant step-by-step escalation.

The second kind of occasion where the two principles reveal themselves is where the speaker supposes that the recipient can identify the referent from a minimal form, but in fact the recipient can't:

(40) (Sacks & Schegloff 1979: 20)
A: Hello
B: 'Lo
Is Shorty there,
A: Ooo jest- Who?
B: Eddy?
Woodward?
[
A: Oo jesta minute

The recipient of a minimal recognitional who cannot recognize the referent from the reference form initiates a try-marked escalation quite parallel to the speaker-initiated kind we have just seen. There are differences, e.g. in the first kind, the recipient asserts recognition of the referent as soon as possible, while in the second kind the recipient upon recognition gets straight on with the business in hand, but both these much replicated sequences serve to indicate that:

(a) there are two potentially conflicting principles at work,
(b) one a principle of minimization, the other of recognition or recipient design, such that
(c) recognition takes precedence over minimization where necessary, but minimization is only relaxed step by step until recognition is achieved,

(d) such conflicts of principle being resolved by precisely replicated sequences like these.

In retrospect, we can now see that the sequence truncation in the case of self-identifications is just this same process of preference for minimal reference to persons in the domain of first person reference. This little fact about English conversational organization might of course be pretty anecdotal and happenchance evidence for two mighty principles of the sort we are trying to establish. It is therefore of considerable interest that some painstaking work by Moerman (1977; n.d.) has shown that, when due allowance is made for different cultural modes of reference to persons and language-specific modes of 'try-marking', more or less identical principles operate in conversational Thai. And an Australian language which will concern us in section 4., Guugu Ymidhirr, can be shown, *mutatis mutandis*, to exhibit exactly the same principles.

Consider:

(41) REV-GEST 107-111

1 B: *‹aa nubuun nhangu waami ( ) milgamul?* ‹ = gesture SW
and I came across the-deaf-one
(.)

2 R: aa ((inconclusive))
huh

3 B: old Tommy Confon

4 R: aa ((nods))
Oh yes

5 B: nyulu galmba duugan duugan dhadaaray bada
he was coming up from Bridge Creek ...

On the videotape, at the beginning of line 1 B brings his gaze round to look at R (they are sitting side by side, and in any case, gaze is normally avoided), to see whether the epithet (a standard minimal reference form for persons) together with his clear gesture to the South West where the referent used to live, will be sufficient to identify the referent. That he has some doubt is shown also by a high rise-fall on the epithet *milgamul*, which I take to be the local equivalent of a 'try marking' intonation. The recipient R produces an *aa* ('huh') that is marked prosodically to convey uncertainty (by low amplitude in particular). B then upgrades the referent with his English name and the descriptive English adjective *old* (such code-switching itself possibly marking this upgrading). This is produced with a rising intonation that I again take to

be 'try-marked'. During this turn he gazes 'inquisitively' at R. R then produces another *aa* with greater amplitude together with a distinct nod. B having received notice of recognition, then withdraws his gaze immediately and proceeds directly with the story (line 5). Allowing for the use of epithets, slightly greater latitude in intonational cues that count as try-markers, gestural specifications of referents[13], and the marked nature of gaze in Guugu Yimidhirr interaction, all the ingredients of the Sacks & Schegloff sequence pattern are here displayed.

### 2.7. *Summary and implications*

We have reviewed a range of CA findings that in various general ways support the view that there are principles of minimization governing conversational sequences. Some of these properties of conversational sequences are in line with what we have called *minimization1*, i.e. with informational constraints: thus the principle that one should not tell people what they already know, and the 'membership categorization devices' for enriching interpretations, are directly parallel to the I-principle (minimization of information and its interpretive corollary, the enrichment rule). On the other hand, the way in which the turn-taking system operates to restrict turn-size, and some of the ways in which repair sequences tend to be truncated, seem more in line with constraints on brevity of expressions than with constraints on information, i.e. with *minimization2*, perhaps properly attributed in a Gricean framework to a maxim of Manner.

However, in most of the reductions considered minimization1 and minimization2 go hand in hand, so that, for example, truncated pre-sequences necessarily involve both kinds of minimization simultaneously (and similarly for most of the truncated sequence patterns described in 2.5).

Where truncated sequences are spelt out, one tends to find *step by step* escalation of informational cues, as in self-identifications, where minimal paralinguistic cues (voice-quality on greetings tokens, etc.) give way to maximal paralinguistic cues ("It's me"), and successively to richer lexical specification ("Les (.) Garston (.) the man who rang about the plumbing"), a pattern also neatly exemplified in references to third persons.

This step-by-step escalation of communicational cues provides a rather different model for the interaction of the I- and Q-principles. In the Gricean framework our solution was a set of rules assigning priority to one principle or the other according to the nature of the lexical cues (viz. if lexical items are drawn from a strong contrast set, Q-inferences predominate over potential

rival I-implicatures, but otherwise I-enrichment happily proceeds unless a marked linguistic form triggers a complementary Q/M implicature). That was a solution proposed without regard to the processual nature of conversation. In the latter dynamic perspective, it is clear that the interaction of two such contrary principles does not have to be resolved within a single clause or utterance-unit: the speaker can try a minmal form and see whether it succeeds, the interactional nature of conversation allowing interlocutors to signal uncertainty about interpretation in the next turn, or even before that by the momentary withholding of next turn, such uncertainty engendering the step-by-step escalations we have reviewed. An important contribution of recent CA work has been to show that mid-turn adjustments of this sort are the rule. Such a processual and interactional solution makes much more comprehensible the interaction of antinomic principles like Q- and I.

Nevertheless, the Hornian 'division of labour' does appear to have analogues in empirical findings about conversational organization. The clearest of these are to be found in preference organization, where preferred responses are short and unmarked, while dispreferred responses are drawn out, marked, and replete with accounts, etc. When we ask what actions are normally performed in preferred (unmarked) format, we find that perhaps the best characterization would be 'the expected, the stereotypical' responses, rather than the responses speakers desire (this is evident from the fact that blamings get denials in preferred, unmarked format; see Atkinson & Drew 1979: 122ff). Thus from minimal unmarked responses to first parts of adjacency pairs, a participant can understand the full stereotypical or expected response; any deviation from the unmarked response will tend to be interpreted as a dispreferred, non-expected, non-stereotypical action forthcoming or being withheld. The sensitivity to these slightly marked forms is very striking indeed, and the negative Q/M-like inferences quite dependable.

One point, though, that does emerge clearly from the CA literature is that these minimizations are not solely motivated either by a Zipfian least-effort principle or simply by a communicational convention that 'unmarked forms convey unmarked meanings'. Rather, one can find, at least in many cases, motivations that derive from the social nature of interaction. One important social principle seems to be: 'if we are socially close we don't need to be explicit'. Another seems to be: 'if there's an element of impropriety, don't say it, implicate it'. These two rather different sources of implicitness were brought together by the theory of 'face'-oriented interaction derived

from ideas of Goffman and developed by Brown and Levinson (1978), but let us take them one by one.

Minimization as an indication of social closeness has strong intuitive bases, and has been much speculated about (Bernstein 1973; Kay 1975; Brown & Levinson 1978). One can see in conversational sequences some reasons for this:

(42) (Drew 1984: 142 [Rahman: B:1,1,12:1]; slightly simplified)
I: Yeh. .h uh:m (0.2) I've jus' rung to to- eh tell you (0.3) uh the *th*ings 'ave arr*i*ved from Barker 'n *Sto*nehou:se
J: [*Oh*::::: (.) Oh can I come round hh
I: Yes please that's what- I wantche to come *rou:*nd
J: [ha hah a:h

In this sequence, a reporting of the arrival of some new furniture presages an invitation to view, but that invitation is pre-empted by a self-invitation. As a result we get a truncated version of the full sequence that might have gone:

I: (preamble reporting)
J: (news receipt)
I: (invitation)
J: (acceptance)

We also have a minimal reference form (*the things* being semantically general). Of course, no such sequence might have taken place if the recipient had treated the report as merely a report. It is *seeing the report as a potential invitation* that is the essence of the 'sociability' displayed (Drew 1984: 142-3).

But seeing something inexplicit as a full-scale invitation is a highly vulnerable assumption:

(43) (Heritage 1984: 319)
B: So::: we thought that y'know=
=if you wanna come on *over* early. Come on *over*
A: .hhh- .hhhh::::: *Ah*:: hh fer *dinner* y' mean? hh
B: No *not* fer dinner. h=
A: =Oh

and it is the vulnerability of either or both interlocutors (in two party conversation, or all in multi-party conversation) that can be seen to motivate many further kinds of minimization. Consider for example pre-requests that check

out whether the conditions for a request are met; if the conditions are not met, or it would clearly be difficult for the addressee to meet them, the request can be aborted. And given that the request is often fairly transparently forthcoming, the addressee is given a chance to invent pre-occupations that would preclude the request. Thus interactional space is provided to negotiate whether the addressee is actually open to a request at all, all without the issue ever becoming explicit. On the other hand, if the addressee is open to compliance, the implicit foreshadowing of a request allows him or her to do the kind of pre-emptive offer that we have seen is of the essence of 'sociability':

(44) (Levinson 1983: 343)
C: Hello I was just ringing up to ask if you were going to Bertrand's party
R: Yes I thought you might be
C: Heh heh
R: Yes would you like a lift?
C: Oh I'd love one

It is this interactional negotiation of delicate social matters 'off the record' as it were that seems to provide one major motivation for minimization (for an attempt to specify what counts as 'delicate' and why, see Brown & Levinson 1978).

We thus return to the notion of a social 'economy' of information and its expression, for which much recent CA work gives ample evidence (see papers in Atkinson & Heritage 1984, especially Part II, and Ch. 9). Do these findings make otiose a maxim of minimization and all the rest of the Neo-Gricean apparatus that we developed in section 1.? Is the principle of minimization at best some generalized conflation over a whole range of independent social motivations for implicitness, and at worst based on an absurd de-socialized view of the rational agent? I do not believe that there is in fact any necessary conflict, for what the Neo-Gricean ideas specify are default *principles of interpretation* that apply to certain alternative modes of expression. In effect, they predict a 'code' of preferred interpretations, as it were; how and why this 'code' is employed in the ways it empirically is, is properly accounted for by social and interactional motivations.

Let me give two examples. Sequence-truncation should, by virtue of the I-principle, be associated with stereotypical and thus preferred responses or activities; thus, on this theory, one would expect sequences of a pre-request

followed directly by a refusal to be (a) rare, and (b) to carry 'violative' implicatures (for an example, see Levinson 1983: 362). But why requests in particular would often be prefaced with pre-requests is something for which we need a different, interactional motivation, in terms of the avoidance of possible improprieties. Again, reconsider (42): the I-principle will predict that a pre-emptive response to a report that might presage an invitation will be interpreted if possible as an invitation acceptance — that's because truncation induces stereotypical interpretation, as just noted. But why preemption in this locus should constitute 'the essence of sociability' is to be accounted for by our other social principle, whereby social closeness is indicated by presumption. Both examples thus seem in line with a division of labour, wherein the Gricean framework predicts preferred interpretations attached to alternative linguistic expressions, while the socially oriented framework explains the deployment of the linguistic resources inducing those interpretations. On this view, the frameworks outlined in sections 1. and 2. of this paper are complementary views of the same phenomena.

## 3. Syntax and minimization

### 3.1. *Syntax and pragmatics: Nature of the interaction*

If this program has anything to recommend it, then one might expect some close interaction with principles of syntax. Before embarking on some sketchy speculations here, it is important to distinguish two rather different possible claims about the interaction of pragmatics and syntax:

a. *Pragmatic reductionism*, which seeks to show that a systematic pattern of distribution or construction is actually due not to a rule of grammar but rather to a preferred rule of use, itself following from a more general usage principle;

b. *Pragmatic complementarism*, or *potential parallelism*, which seeks to show that such a systematic pattern, which may or may not be specified by a rule of grammar, is consistent with a pragmatic principle specifying use.

The strong reductionist claim is a view foisted on us by the generativist assumption of explanatory parsimony: *for each fact there should be one and only one explanation*. So generative grammarians, as soon as they are convinced that there *is* a functional or pragmatic explanation of syntactic facts, tend to assume immediately that the facts are therefore *eo ipso* no longer syntactic (see e.g. Lightfoot 1979: 43-4). This move may possibly be justified if one is involved in systematic theory construction in a limited domain, but it

is of course absurd as an ontological claim — the world is full of multiple constraints; and when one is considering the interaction of two rather different domains, like syntax and pragmatics, it is absolutely the wrong model of the interaction between them.

The second, weaker claim, of potentially overlapping explanations is implicitly assumed by all the work that goes under the 'from discourse to syntax' rubric (e.g. Brown & Sankoff 1976; Givón 1979), wherein it is presumed (a) that functional pressures remain constant, and (b) that solutions to functional pressures tend to get syntactized. It follows from those presumptions that there is every reason to expect co-existing, complementary pragmatic and syntactic explanations of the same distributional patterns.

Of course, in this second framework, it is important to keep the diachronic and synchronic perspectives analytically distinct: an existing syntactic construction may have clear origins in a usage pattern, explained by pragmatic principles, which is its diachronic source. With the diachronic process of grammaticalization goes of course a process of semanticization: the pragmatic inferences associated with a usage pattern lose their defeasibility and become the semantic implications of a grammatical construction. Thus it will be argued below that the coreferential zero-anaphor in *I want O to go* conforms to the general pragmatic principle inducing I-implicatures of coreference from minimal reference forms, even though in this case such a pragmatic account is otiose, coreference (or at least overlapping reference) being grammatically encoded. The conformity is still of considerable interest in the complementarist perspective, as I hope to show.

### 3.2. *English anaphora*

The area we shall dabble in is the interaction of two kinds of inference and their relation to syntax, namely: (i) the inference of conceptual relations between clauses; (ii) the inference of co-reference between referring expressions. These are often treated as totally independent issues, but it is quite clear that the inference of one will aid and abet the inference of the other.

#### *3.2.1. The informative interpretation of pronouns*

One functional way of thinking about inter-clausal syntax is that clause-linkage devices exist to combine clauses for just two main reasons: (a) to compress them — to minimize the length of expressions; (b) to constrain interpretation — to reduce the hearer's search domain for kinds of semantic linkage between the clauses.

Given the general direction of the I-principle, we can expect that com-

pression will tend to lead to stereotypical readings. Consider for example:

(45) a. John turned the switch and he started the motor
b. John turned the switch and Ø started the motor

(Incidentally, throughout the rest of this paper I shall mark NP-gaps as in (45)b. with a zero, calling the phenomena 'zero-anaphora', but without any theoretical commitment to the existence of abstract entities, or phonetically-null NPs.) Both of these sentences, given the I-induced process of 'conjunction buttressing' described in 1.1, will tend to I-implicate non-truth-functional connections between the two conjoined clauses, especially the following relations in the following order (as far as is consistent with what is taken for granted):

(46) close temporal sequence > cause > intended result

But the reduced form in (45)b. reinforces the stereotypical interpretations in (46). (On the pragmatics of conjunction-reduction see Schmerling 1975, Harnish 1976.)

Let us now consider reference to persons and the inference of co-referential relations between referring expressions. We have the options of descriptions and names; but we also have provided minimal reference forms — pronouns and the ultimate minimization, the zero-anaphor. Let us just concentrate on the options:

(47) Full NPs > pronouns > Ø

It has often been noted that the minimal forms tend to be used for 'given' referents, especially after first introduction by full NP (see e.g. Foley & Van Valin 1983: 327). So the coreferential reading of (45)a. is canonical. But the inference of co-reference between the full NP and the pronoun is merely a favoured reading: *John* and *he* could be disjoint in reference. Nevertheless, we must account for the favoured reading.

But before showing that there is an independent I-implicature of co-reference between the full NP and the pronoun, we should note that the inference of inter-clausal linkage will tend to give us exactly the same result. For, given the tendency to 'conjunction buttressing', there will be a tendency to assume that *he* and *John* are coreferential. This is because in seeking close conceptual relations between the two clauses, we will note that if we have the same agent involved in two events then we must assume consecutive actions, and if we have the same agent in two events then that lends itself especially to a teleological connection. And in going for a teleological connection, we

are after the closest possible linkage between the two clauses, namely the interpretation under which they are *one* action under a complex description.

Now let us turn to the inference of coreferentiality independent of the clause-linkage inference. The pronoun *he* is definite but semantically general: all that is specified is a male animate entity. There is no Q-inference from a scale or contrast-set to halt an I-inference, which therefore goes through. The I-principle will promote a more specific interpretation; given that *he* is definite, if there is no other competing referent in the immediate domain of discourse, the 'best interpretation' will be the highly specific reference picked out by *John*. Hence we can assume co-reference. Of course, the zero-form in (45)b. is even more semantically general, there being no restrictions on number or gender, and the arguments here will go through for the zero-form too — in this case redundantly, since, I assume, coreference is here stipulated by a rule of grammar or semantic interpretation (but the pattern is important as in many languages the interpretation is not so stipulated, see 4. and Levinson n.d.).

The claim that the I-principle will naturally induce co-referential readings of pronouns (and other reduced NPs) has of course far-reaching consequences, and would need more justification than I have space for here (see Levinson n.d.). Suffice it to say that there are a number of arguments that converge to make plausible the idea that co-referential readings are more informative than disjoint ones. The first line of argument is at the level of reference: the smaller the domain of discourse, the more informative the statement (since the number of compatible states of affairs will be reduced, see Bar-Hillel & Carnap 1952); also, the postulation of extra entities in the domain of discourse weakens the strength of a statement, due to the relative unfalsifiability of existential commitments (Popper 1959: 68ff). The second line of argument is at the level of sense: clearly, (45)a. will be more informative *about John* where *he* is taken coreferentially with *John* than it would be about either referent where two are posited and *he* and *John* are disjoint. I assume therefore that we can indeed provide a derivation of the preferred co-referential interpretation of pronouns from the I-principle, which induces maximally informative readings from semantically general or otherwise minimal reference forms.[14]

Notice, then, that this I-implicature to coreference from the pronoun, a semantically 'reduced' form in the sense of minimization1, could be defeated by a Q/M inference from a more marked form, as predicted by the Hornian 'division of labour':

(48) John turned the switch and *the man* started the engine

Without the Hornian counter-balancing Q/M implicature, the I-principle would of course make the wrong predictions here, since *the man* is semantically general and ripe for I-enrichment.

So, to sum up so far: the inference of clause-linkage, i.e. 'conjunction buttressing' by the I-principle reinforced by the reduced form, combines with an independently derived I-inference from the general term to the specific interpretation, to yield a strongly favoured reading of co-reference between the full NP and the pronoun.

*3.2.2. Government and Binding Theory and Q- vs. I-implicatures*

(a.) *Conjoint and disjoint readings of pronouns*

However, as long noted, things are not as simple as this. Full nouns and pronouns often have disjoint (non-coreferential) readings as in (49) (square brackets indicate relevant 'Minimal Governing Categories'; 'indices' indicate normal co-referential or non-co-referential interpretations):

(49) a. [Joan1 adores her2]
b. She1 doesn't like the dress [Joan2 bought her1/3]
c. [Joan1 adores herself1]
d. She1/2 doesn't like the dress [she1 bought herself1]

All this is only too familiar from Government and Binding (GB) Theory (Chomsky 1981), and it's usually been stipulated that pronouns have to be free in their minimal governing category — that is, not coindexed with a c-commanding NP in their minimal governing categories (MGCs). Pronouns in this respect contrast with both lexical NPs and reflexives/reciprocals, as stated in the three Binding Conditions in (50):

(50) A: 'Anaphors' (reflexives and reciprocals) must be bound in their minimal governing category;
B: Pronouns must be free in their minimal governing category;
C: Lexical NPs must be free everywhere.

But Reinhart (1983) has recently proposed that in fact not all three conditions in (50) are grammatically specified; rather, the conditions on pronouns and full NPs follow by pragmatic inference from the condition on anaphors and quantifier-bound pronouns, which remains a rule of grammar. The argument is interesting because it relies crucially on precisely the kind of pragmatic principles we have developed, and whether or not it is correct it illustrates

a general possibility, namely, wherever there is a grammatically restricted interpretive procedure there is a pragmatically generated mirror-image inference.

Suppose, she suggests, we have just one syntactic indexing rule, permitting coindexing in just two conditions (my restatement of her rule, 1983: 158) as in (51):

(51) *Reinhart's Coindexing Rule*
Co-indexing permitted between italicized expressions only when:
1. A *reflexive/reciprocal* is c-commanded by an *NP* within its minimal governing category
2. An ordinary *pronoun* is c-commanded by an *NP* outside its minimal governing category

This rule is optional, but non-coindexed reflexives will be filtered out as uninterpretable by a separate device (1983: 159).

Note that the second condition in her rule in (51) allows co-indexing of a pronoun and a lexical NP only if the latter c-commands it (very roughly, is higher in the tree); this is to distinguish the different coreferential possibilities as illustrated below:

(52) a. Zelda1 loves her1/2 neighbours (coreference possible)
b. She1 loves Zelda's2 neighbours (coreference impossible)

We are now in a position to state Reinhart's pragmatic argument to explain all the preferred conjoint/disjoint interpretations that are not covered by her syntactic and semantic rule. She proposes a maxim of Manner "Be as explicit as the conditions permit" (1983: 167). Then she can make the argument that failure to be explicit pragmatically induces a disjoint reading:

(53) a. Joan adores her
b. Joan adores herself

If the speaker meant (53)b., he shouldn't, given the above maxim, say the weaker (53)a. (where 'her' could refer to anyone under the sun, including Joan); therefore if he says the (53)a. sentence, he implicates that the sentence in (53)b. does not fit the facts. This is of course in fact an application of our Q-principle, and not due to a maxim of Manner at all, for it is an inference from the use of an *informationally* weaker expression to the non-applicability of the stronger expression (the reflexive). There is a Horn-scale or contrast set ‹herself, her›, so that the use of the weaker non-reflexive pronoun Q-implicates the negation of the stronger reflexive anaphor.

Systematically applied, such reasoning will generate the complementary distribution of 'anaphors' (reflexives and reciprocals) and pronouns on a given indexing which is at the heart of GB theory — indeed it amounts to a partial pragmatic reduction of that theory (Levinson n.d.). For the distribution described in Binding Condition B of (50) is predicted by the Q-principle applied to the distribution grammatically specified by Binding Condition A: pronouns will be (by Q-implicature) disjoint in reference wherever a reflexive could have been used to encode coreference. In the same way, lexical NPs will be read as disjoint wherever a reflexive could have been used to encode coreference; thus part of Condition C is also given by the Q-principle applied to Condition A. Where pronouns can not be replaced by reflexives (for example, in subject position or where they are genitive modifiers as in *his book*), there will be no Q-implicature to block I-enrichment from the reduced pronominal form to a preferred coreferential interpretation. However, in those same locations where a reflexive could not have been used and a pronoun would tend to implicate co-reference, a full lexical NP will Q/M-implicate disjoint reference by virtue of its marked form compared to a pronoun — thus obtaining the remainder of the pattern described in Condition C.

Less straightforward is to explain the difference between the coreferential possibilities in (52), which seem to depend on the position of the pronoun in a syntactic configuration. Reinhart's argument is that her grammatical coindexing rule applies optionally in structures like (52)a. but cannot apply to the sentence in (52)b., due to the requirement in rule (51)(2) that the pronoun be in a c-commanded syntactic position. Although the rule only optionally applies to (52)a., the fact that a sentence of this type potentially encodes coreferentiality makes it the proper way to express the conjoint reading, by virtue of the Q-principle ("say as much as is required"). So if, given the possibility of saying the informationally richer statement that can grammatically encode co-reference, the speaker uses the weaker form as in (52)b., even though *grammatically* reference is free, there will be a strong Q-implicature to the negation of the stronger reading, i.e. an inference to the disjoint interpretation of (52)b.

The advantage of the Q-implicature account is that, implicatures being defeasible, it can accommodate the marked exceptional usages that embarass the GB account. For example, take the Binding Condition C prediction that the lexical NPs in *Felix hit Felix* or *Felix hit the man* will be disjoint in reference. On our account this is due, not to a rule of grammar, but to a Q-implicature from the fact that the speaker has failed to use the stronger *Felix hit*

*himself* which would grammatically encode coreference. But Reinhart points out that (54)a. is in fact quite acceptable on the coreferential reading indicated by the 'indices', because the normal alternative using a reflexive, as in (54)b., would in fact here mean something different. (To see the difference in meaning between (54)a. and b., consider that (54)a. entails that Felix only got one vote, but there is no such entailment of (54)b.) Thus, in this case, the reflexive and the lexical NP are not in contrast, and no Q-implicature arises from the non-use of the reflexive to block the assumption of co-reference.

(54) a. Only Felix1 voted for Felix1
b. Only Felix1 voted for himself1
c. The vendor1 reserves all rights to the vendor1

And we may also note here that there are registers and speech events where sentences like (54)c., with the interpretation indicated by the indices, are quite routine — suggesting a genre-specificity to the way that the Q- and I-principles apply, a matter to which we'll return.

Reinhart's theory thus uses the Q-principle to do away with a rule of grammar specifying disjoint reference between lexical NPs and pronouns, and in doing so gains the elasticity afforded by the defeasibility of implicatures. Her sketchy pragmatic account seems to be thoroughly fleshed out by bringing to bear the systematic interaction of the Q- and I-principles as sketched in the resolution schema in (23).

(b) *PRO: Zero-anaphora and clause linkage in English*

Let us now come back to the opposition between pronoun and zero form. This opposition, as it shows up in the study of English infinitival complements, has been the subject of a great deal of thought. In Government and Binding Theory, there is of course a special sub-theory, the theory of *control*, to handle the interpretation of the zero-subject (designated PRO) of such infinitives. The classic problems concern the distinction between subject-controlling verbs as in (55), and object-controlling verbs as in (56):

(55) (where PRO = Max = Subject of Matrix)
a. Max tried [PRO to eat acorns]
b. Max promised Sue [PRO to read the book]
c. Max asked Mary [what PRO to do]

(56) (where PRO = Sue = Object of Matrix)
a. Max persuaded Sue [PRO to go]
b. Max forced Sue [PRO to go]
c. Max reminded Sue [PRO to go]

Jackendoff (1972), Radford (1981) and latterly Foley & Van Valin (1983) have argued that these patterns are determined by the semantics of the verb. Thus Foley & Van Valin point out that Subject-Control verbs are volitional or report commissive speech acts where, naturally enough, the agent (normally matrix subject) is to do the infinitival action; while the Object-Control verbs are causatives or report directive speech acts (a kind of causative), where naturally the causee (normally matrix object) is to do the infinitival action.

However, there are some indications that this is not perhaps so firmly grammaticalized; consider the following well-known cases where the interpretation varies not according to the verb but according to the most probable scenario:

(57) a. Zelda asked Mary [PRO to leave]
b. Zelda asked [PRO to leave]
c. John appealed to Bill [PRO to leave]
d. John appealed to Bill [PRO to be allowed to leave]

In English, this flexibility of interpretation is not great, but in other languages it can be extensive (see Comrie 1985, on German and Russian; Levinson n.d. on Guugu Yimidhirr). Thus it is worth rehearsing an account of control that assumes that control is pragmatically determined withhin semantic limitations, acknowledging that in English such an account may be largely preempted by grammatical stipulation of control (recollect our assumption of possibly parallel syntactic and pragmatic explanations). It might go as follows:

1. In all cases where only one actor is mentioned (whether or not this is syntactically constrained as with *try* or merely possible as with *ask* in (57)b.), the addressee may assume the same actor is the agent of the infinitival clause. (Note the contrast: *John1 asked PRO1 to go* vs. *John asked Mary2 PRO2 to go*.) Such an assumption will be in accord with the I-induced search for 'the best interpretation', for the alternative would be an indefinite interpretation that would be less informative. Notice here that the use of the PRO-infinitive construction is a *reduced*, compressed or minimized version of the (sometimes parallel) *that*-S construction, as in the pair (i) *John promised PRO to come,* (ii) *John promised that he would come*. In (ii), *he* is, as we have seen, not grammatically specified in reference, but there is an I-inference to the specific coreferential interpretation. So, exactly as usual, the reduced form in (i) will favour the most informative interpretation, i.e. it will reinforce, and make less defeasible, the inference to coreference.

2. Where two actors are mentioned in the matrix, the addressee may prefer the most likely interpretation of coreference bettween one of the referents and PRO, given the causative-directive vs. volitional-commissive predicate. Thus in (57)d. the subject-control interpretation may be due to the sheer conceptual difficulty of imagining a scenario in which the object-control interpretation (as in (57)c.) would make sense (John would be appealing to Bill for Bill to permit himself to leave). In essence, then, in the case of matrix verbs with two arguments, the interpretations would simply be inferences to the 'best interpretation' given the semantics of the verbs and the likelihood of certain scenarios. A pragmatic explanation would explain the possible reversal of the normal coreference assignments, as illustrated in (57) above.

Let us now turn to just certain structures where PRO is in systematic opposition to an overt pronoun or lexical NP. Given GB theory, where PRO must not be governed and overt NPs must be governed in order to obtain case, the two will tend quite clearly to be in complementary distribution. There are just certain exceptional loci, as in the following:

(58) a. John wants PRO/him to come
b. John is keen PRO/for him to come
c. John likes PRO/his going away

So, in just these locations we have a possible contrast, PRO (or NP-gap) vs. overt pronoun. And we would expect the use of the more prolix overt pronoun to Q/M-implicate a referent disjoint from the PRO interpretation induced by the I-principle. And, as Chomsky notes, this is exactly what happens (indices here mark preferred interpretations):

(59) a. He'd1 prefer 01 going
b. He'd1 prefer his2 going (favoured interpretation)
(60) a. He1 wants [PRO1 to go]
b. He1 wants [(for) him2 to go]
(61) a. John1 bought a book [PRO1 to give to Mary]
b. John1 bought a book [for him2 to give to Mary]

To account for these interpretations Chomsky invokes an '*Avoid Pronoun' Principle* (1981: 65), which specifies "a choice of PRO over an overt pronoun where possible" — i.e. in structures where either PRO or pronoun may appear, the use of the pronoun implies disjoint reference. He adds that this Principle "... might be regarded as a subcase of a conversational principle of not saying more than is required, ... but there is some reason to believe it functions as a principle of grammar". Our methodological parallelism would

allow both, and as Horn points out the Avoid Pronoun Principle is entirely consonant with the I-principle (1985: 23) and his 'division of labour' between I and Q/M (1985: 25). Moreover, we can now be a little more precise using our revised schema for I-, Q- and Q/M-principle interaction:

1. There is no Horn-scale ‹pronoun, PRO› to prevent an I-implicature enriching the reduced form;[15]
2. The I-principle will induce a maximally informative interpretation of the reduced (PRO) structures, yielding the assumption of co-reference wherever this is consistent with what is taken for granted;
3. Where, in contrast, a more prolix form (the pronoun) is employed, this will Q/M-implicate that the stronger conjoint reference is not applicable.

I shall argue below (4.) that these patterns of interpretation of zero- and contrastive forms can be observed to operate in an identical way in a language where subjects (and other NPs) are freely omissable, and that in such a language they play a crucial role in the understanding of discourse.

Clearly, in English there are tight constraints on zero-anaphora; it occurs more or less only in subject position[16], and then only in the sorts of structures we have been reviewing. These structures are quite clearly matters of interclausal linkages of various kinds, and express close conceptual relations across clauses (e.g. connections between a propositional attitude and the proposition entertained; or between verbs of saying and the matter expressed; etc.). It is these tightly knit clause-linkages that permit zero-anaphora in English, and the zero-forms themselves reinforce the highly informative interpretations, as we have seen.

This particular relation between minimization and informativeness has been noted to hold on a cross-linguistic scale. Building closely on work of Silverstein's (1976), Foley and Van Valin (1983) have suggested a hierarchy of syntactic bonding between two clauses. This hierarchy is expressed in a theory-dependent way in terms of three sizes of unit (roughly, verb, verb plus subcategorized arguments, whole clause including adjuncts) and three kinds of bonding between the units (coordination, subordination and 'co-subordination'), yielding nine degrees of bondedness so that two verbs co-subordinated are most tightly linked and two whole clauses co-ordinated are least tightly linked (1983: 267). To this hierarchy of syntactic bonding there is closely correlated a hierarchy of semantic bonding (1983: 269; cf. Silverstein 1976: 163), roughly as in (62):

(62) *Semantic relations hierarchy* (Foley & Van Valin 1983: 269)

| | |
|---|---|
| *Strongest* | Causative |
| | Modality (e.g. 'try to') |
| | Psych-action (e.g. 'want to') |
| | Jussive (e.g. 'order to') |
| | Direct perception complements (e.g. 'see S') |
| | Indirect discourse complements (e.g. 'say that S') |
| | Temporal adverbial clauses (e.g. 'Before S, S') |
| | Conditionals (e.g. 'If S, S') |
| | Simultaneous actions (e.g. 'While S, S') |
| | Sequential actions - overlapping |
| | - non-overlapping |
| *Weakest* | Action-Action (unspecified linkage) |

The authors claim that the syntactic hierarchy and the semantic hierarchy are closely aligned in all languages, so that the greatest morpho-syntactic adjustments are used to express the tightest semantic linkage. However, the relevance of all this to us is the following claim: "the tighter the linkage, the more likely the coreferential argument [shared between two clauses] will be expressed as zero. ... Exactly where on the IRH [Interclausal Relations Hierarchy] a language will start using zero anaphora will vary, but it will always be the case that the linkage types in which zero anaphora is used will be tighter (i.e. higher on the IRH) than those in which it is not" (1983: 366).

What these suggestions amount to is exactly the same finding that is at the heart of the current proposal: *reduced forms, and zero-anaphors in particular, tend to be associated with the most informative interpretations*. One important feature that the two accounts share is the association of the conceptualization of clause linkage with the determination of the reference of zero-forms. It is striking that research tunneling from the other end, as it were, i.e. from the study of the grammatical encoding of semantic relations, would arrive at the very same place to which the study of patterns of language usage has independently led us.

## 4. Minimization in Guugu Yimidhirr

I want now to turn briefly to consider the interaction of our I-principle and the syntax and discourse structure of an Australian language, Guugu Yimidhirr. Guugu Yimidhirr (henceforth GY) is spoken in the far North of Queensland, Cape York, and is the first language of a group of about 600

aboriginal people who live together on a large reservation. It is undergoing intensive study by John Haviland (1979, n.d.), but I've also had the opportunity to do some fieldwork there and record and film stories and conversation. The following remarks build directly on Haviland's thorough grammatical work[17], although the data and analysis here are my own.

GY is a non-configurational language in Chomsky's (1981) terms and, more specifically, a classic W* language in Hale's (1983) terms. There is an extraordinarily free word order (made possible by elaborate case marking — GY is a split ergative language); not only can elements of something one would ordinarily call an NP be distributed over the length of a clause, even clauses can sometimes get jumbled up together (see e.g. Haviland 1979: 154ff on 'ergative hopping'). In addition, under the right discourse conditions, there is free 'deletion' (i.e. non-appearance) of the valents or arguments of a verb.

In addition to this, GY has very little interclausal syntax: the only kinds of syntactized clause-linkages are *causal* clauses (translating sentences like *Because John came, Bill went*), *purposive* clauses (translating sentences like *John went in order to fish*), and *simultaneous* clauses (translating sentences like *While Tulo slept, Roger ate*). Notably lacking are English subordinating constructions like adverbial clauses of time and place, relative clauses, and co-ordinating constructions like explicit conjunctions. All these concepts have to be signalled by parataxis between two independent clauses.

A final important feature of GY syntax is that it falls within the fourth class of the Foley & Van Valin taxonomy of reference-tracking systems (1983: Ch.7; Van Valin, this volume); it has what they call an *inference system*. The whole 'problem' of reference-tracking is actually a direct product of our minimization principle: given the fact that full descriptive or proper name NPs tend to get reduced to pronouns or even zero-anaphors, the problem arises how in a narrative one is to guess who is doing what to whom. The solutions, suggest Foley and Van Valin, vary from marking the gender class of nouns on the verb, to indicating switch of main protagonist on the verb ('switch reference systems'), to keeping main protagonists as subjects (even when not in agent role) by the use of the passive, as in English ('switch function systems'). Finally, we have 'inference systems', that is, deletion without syntactic clues for recovery, i.e. no grammatical solution at all. GY exhibits an inference system because while allowing extraordinarily free deletion of NPs, there is no cross-indexing on the verb that would allow one to infer by 'gender' class, or switch of reference, or switch of function, what has been

deleted.

Just how free deletion in GY is can be gleaned from the fact that the only clear deletion constraints in the tightest GY structures, namely those three kinds of subordination, are, roughly, that the agents of transitive subordinate clauses can't (or at least preferably don't) delete under identity with the objects of matrix clauses (there are exceptions even to this; see Haviland 1979: 135, Levinson n.d.).[18]

There are thus at least three bundles of syntactic properties that lend some very special characteristics to GY texts:

(i) the syntactic flexibility of the simple sentence, allowing missing valents;

(ii) the relative lack of explicit syntactic marking of semantic relations between clauses; and finally:

(iii) a 'pragmatic' reference-tracking system, with zero-anaphora permitted in nearly all syntactic positions without any compensating verbal cross-referencing.

These syntactic features allow texts to be constructed where there have to be massive amounts of I-implicatures. If one could imagine a typology of languages that placed them on a scale between an I-pole (with minimal specification requiring maximal inference) and a Q-pole (with maximal specification and minimal inference), then Guugu Yimidhirr is the paradigmatic I-language! The best way to get a feel for this is to look at a sample text.

The following are extracts from a story about events that happened over 50 years ago while the two narrators[19] were boys and many of the aboriginal people still roamed the bush. The story is about a boy (participant B's classificatory 'nephew') who drowned, and how this death was attributed to the supernatural powers of one Douglas, who was therefore murdered by three of B's relatives in revenge. The narration was filmed, which turns out to be important, as gestural information is essential to the understanding of the text.

In the first extract, the three revengers are attacking Douglas. The thing of interest is how the four protagonists are distinguished despite the use of minimal forms. For expositional reasons, let us start by looking at an English gloss of the passage, where zero-anaphors are marked O, all other referring expressions are also italicized, each with a referential index to indicate identity of reference (as it was indicated to me by informants). I've assigned index 1 to Douglas, the victim, 2 to Buli Buli, one of his assailants, and 3 & 4 to the other two assailants. The full GY text will be found in Appendix 1, to which the following line numbers coincide (hence in the English gloss there

are sometimes some missing line numbers, where pauses etc. have been omitted, or two English lines have been assigned to one GY line). It is important to remember that although the English gloss encodes in the verb inflection the singularity vs. plurality of zero-referents, such information is not encoded in the GY text. I should also add that GY pronouns are unmarked for gender, although they are marked for case and number (singular, dual, plural) and presuppose animacy.

(63) (Reveng1.mrg)

167 B: *They*1/2/3/4 looked at one another
168 *O*1/2/3/4 grabbed each other
169 *He*1 was you know-
170 *That chap*1 *he*1 was very big
171 R: *0*1 was very big indeed
172 *He*1 was Jimmy Lee's younger-brother
173 B: Yeah. *O*1 grabbee *another* (e.g. 3)
174 *O*1 threw *O*3
175 R: *O*1 threw *another*3, eh?
176 B: *O*1 threw *another*4
177 *Buli-Buli*2 was very- er
179 R: (*02* was) 'helpless'
180 B: (*02*) helpless see
182 *01* took *him*2 (=Buli-Buli)
183 *01* threw *him*2
184 While *01* was hitting *him that one*3/4 ((points N)
185 *He*2 *this chap*2 to the S came to ((points S))
187 R: *02* got up ((nods S))
188 B: *02* got up and *O*2 grabbed a gun
*02* shot *01* with a gun
191 Bang. The end!
192 R: *01* was finished
193 B: *O*2/3/4 gathered around
*Buli-Buli*2 ifted a rock ((points N))
194 R: *02* crushed *01*'s chest (with it)
195 B: ((bangs chest))
197 *02*/3/4 lifted *01*
198 *02*/3/4 carried *01* West

It will be immediately evident that there is a great deal of zero-anaphora and

a correspondingly large amount of inference required to assign reference to grammatical functions or implicit verb valents. Also, there does not appear to be any simple rule for such assignment. Note for example that successive zero-subjects or topics can be assigned disjoint reference (cf. lines 180 & 182; 194 & 197 etc.). Consider too a possible rule of the sort 'zeros NPs that would be assigned the same grammatical case if they had appeared as overt NPs, get assigned the same referent'; this might account for the coreferentiality of 01 in the clause after 188 and the 01 of 192 (because both, if overt, would be in Absolutive case, as the object of a transitive and the subject of an intransitive clause, respectively, in this Ergative language). However, a counter-example to that possible rule is provided by the preceding first clause in 188: '02 got up' has a zero valent in an Absolutive case role, while in the next clause '02 shot 01 with a gun' the same referent is picked out by a zero in an Ergative case role. Nor does there seem to be any other kind of rule that could specify the reference of successive zeros without essential reliance on complex pragmatic inference.

We therefore need an explanation of how it is at all possible for GY speakers to keep track of referents through a typical narrative passage like this. The following account applies the interaction of I-, Q- and Q/M-implicatures to good effect.

At the beginning of the extract, the pronoun 'they' and the O must contextually refer to the three assailants and the giant Douglas, and by complementarity and singularity 'he' refers to Douglas. In line 171 the zero continues the reference to Douglas, by I-inference from a reduced form to the strongest interpretation. The reversion to the pronoun in 172 is necessitated, I think, by the grammar of identity statements in GY, which requires overt arguments.[20] Then in 173 there is the use of a marked reference proform, *yindu*, 'the other(s)'; unspecified for number, this can't refer to Douglas, because it is a marked form Q/M implicating a *change* of reference, and we've just been talking about Douglas. But this marked form is also in contrast to the pronoun glossed 'they'; thus by opposition to 'they' this form (glossed 'the others') Q/M-implicates a successive division of the three assailants into two parties, the compositions of which are made clear in line 177 by naming the odd man out, viz. 'Buli-Buli'. 'Buli-Buli' turns out to be 'helpless', in fact lame (179-180), explaining his marginal participation in the wrestling. So 'the others' can be indexed as assailants 3 and 4.

Now notice we've already had more zero-anaphors: 'throw' in 174-176 is transitive, so there is a missing valent, the subject (and in 174 a O object I-

linked to the prior object; since we've just been talking about Douglas as 'he1', that referent will make for the closest clause linkages; so these zeros by I-implicature we can index with 1.

In 179 we might seem to have another zero-valent, missing subject for an adjectival predicate; here since we have just named Buli-Buli, the most informative reading is that this zero coindexes with Buli-Buli, hence 02. Actually, a better analysis here must be that 177 displays that B is having trouble with a 'word-search', so that in 179 R is merely supplying the problem word, the predicate alone, to complete B's utterance. (The two analyses make clear the special difficulties of distinguishing between full sentences and elliptical sentence fragments in a language with zero-anaphora.) In any case, we've introduced two focal characters, Douglas and Buli-Buli; so when in 182 the transitive verb 'take' appears with minimal valents, a zero and a pronoun, we know these two referents are subject and object.[21] But which is which? Here an I-inference to the best interpretation in the light of background information seems to be essential: we've established that Buli-Buli is helpless, so if '0 took him', it's probably the giant Douglas lifting Buli-Buli into the air to dash him to the ground, which is indeed indicated by iconic gesture exactly at this point. Once we have that reading, the next line can be assumed to have the same agent and patient, because otherwise we'd expect a marked form, e.g. at least an overt pronoun rather than the zero in agent role, to Q/M-implicate the complement of the I-implicature. We do indeed get an overt 'that one' as object in the next line, 184, so this Q-implicates (by contrast to 'him' in particular, and by Q/M contrast to zero) that 'that' refers to someone other than Buli-Buli or Douglas, which only leaves the other two assailants, index numbers 3 and 4. Again, the subject must be Douglas, because that's the informative reading promoted by the I-principle and there's no marked form to indicate the contrary.

At this point, an entire ancillary mechanism for referential disambiguation comes into play. Guugu Yimidhirr speakers are oriented to the points of the compass at all times, and they make frequent verbal and gestural reference to those cardinal points in order to indicate not only direction of location and motion, but also to pick out referents (as discovered by Haviland 1979, n.d.; see also Levinson 1984). The gestures accompanying this particular narrative are indicated on the GY transcript in the Appendix. Here at 184 we have the details of the event being located in this cardinal point space: '01 was hitting him that one3/4' (with a gesture to the North); and in 185 we are told that 'he came to' (regained consciousness, lit. 'became good') 'to the South'

(with a gesture to the South). The 'he' in 185 must be Buli-Buli; first, there is a full pronoun, Q/M- & Q-implicating 'not any of the ones we've just been talking about', i.e. not 01 or him3/4, so by elimination referent number 2, Buli-Buli. Also, by spatial geometry, referents 1, 3 & 4 are fighting in the North, leaving only 2 as potentially elsewhere, like in the South. In line 187, the other participant R contributes to the narrative, interjecting '0 got up'. Well, so far the zero-subject has been mostly Douglas. But since we have just had Buli-Buli as referent of the full pronoun in the preceding sentence, he's at least as good a candidate for the referent of this zero form, especially as 'getting up' could be a natural sequential action after 'coming to', so that an I-implicature would seek to link these clauses temporally and causally. However, to clinch matters, R at this point gives a clear nod to the South, that is to the location of Buli-Buli in the narrative space we've just set up. So Buli-Buli gets up, with a gun, we are told in 188. In the next line, we can assume (given the I-principle) the same subject, and from the whole direction of the story we can guess the object of 'shoot'; so Buli-Buli shot Douglas (an I-inference to the best interpretation). Notice how in line 192 the zero subject reverts to Douglas, indicating that doubt might really have arisen at 187 without the gestural specification of the referent. Finally, the referents of the zero valents in lines 193 to 198 are given by I-induced inference to the best (most plausible) interpretation, but the actions, and thereby indirectly the actors, are gesturally specified by both iconic and directional gestures.

The role of gesture in reference tracking in GY is very extensive and its frequent occurrence with minimal reference forms (including, especially, NP-gaps) suggests that displacement of communication out of the verbal channel into the kinesic channel is part of the same minimization of verbal means motivated by the I-principle.

Clearly, how the referents in a passage of this sort are successfully tracked is a matter of great complexity, no doubt involving principles beyond those that are focal here. Nevertheless, there are some quite clear patterns to which our three kinds of inference, I-, Q- and Q/M-based, appear to give the key. On the one hand, we have I-inferences from minimal forms that seek close clause-linkages, these often giving us 'same agent/patient as last clause' inferences, or alternatively, suggesting that the best interpretation in accord with what is taken for granted lies in another referential linkage. On the other hand, we have inferences to disjoint reference, triggered either by a contrasting (yet logically compatible) term of the same type (e.g. a pronoun 'they' vs. 'he') Q-implicating non-coreferentiality, or by a form that is marked with

respect to a prior term (e.g. a pronoun vs. a zero; or a marked type of pronoun vs. an unmarked pronoun) Q/M-implicating non-coreferentiality. The latter, Q- and Q/M-inferences, as always, taking precedence over the former I-induced tendency to look for coreference.

In order to elaborate some of these issues, I want now to turn to a later extract from the same story where a pronominal and a zero reference turns out to be actually ambiguous for the participants; showing how the ambiguity arises and how it is resolved is quite revealing about how reference tracking proceeds. In seeking to understand a problem that is also a problem for participants, I attempt to emulate a highly successful CA methodological strategy (see especially Schegloff 1984). After the murder narrated in the prior extract, the local headman King Jacko hears of it. He notifies the white policemen forty miles or so South. The troopers come up North brought by old Tracker Monday:

(64)

266 B: *O*5 (Tracker Monday) also came (with the police)
267 *O*7 (the police) tooke *them*2/3/4. *O*7 took *O*2/3/4 "Show (us) where *you*2/3/4 buried *O*1"
268 R: *0* took *them* back that way ((gestures S)) ←
269 B: That way ((gestures from NE to SW))
270 R: (Ah) that way West ((gestures W))

The ambiguity of interest lies in line 268 (I shall use the term 'ambiguity' in a lay sense to mean 'more than one interpretation'): given 'zero took them back that way', the question is — *who* took *whom*? In the previous lines we've set up Tracker Monday and the police troopers as the joint subjects or agents, with a prior full lexical description and zero thereafter (I've assigned them indices 5 and 7 respectively).[22] From the previous narrative it's clear that the pronoun 'them' and the zero object in 267 refers to the four criminals, collectively indexed 2/3/4.

Given this, the zero subject in 268 could be the same as the zero object in 267, viz. the four criminals, and the pronominal 'them' could refer to the police, i.e.:

(65) Interpretation 1 for line 268
(a) *0*2/3/4 took *them*7 back that way ((gestures South))
(b) This clause is related to the prior ones as a more detailed specification of what the prior ones outlined.

That interpretation may in fact be the one that B presumes, as we shall see. In favour of it, we have not only the standard kind of I-inference from minimal form to co-reference with prior minimal forms (a pronoun and a zero), but also an I-inference to a perfectly plausible clause-linkage: on this interpretation this utterance is linked to the prior ones as an *understanding check*, a request for confirmation of the more detailed specification this 'follow on' utterance proposes. Also in favour of this interpretation is that the overt pronoun 'them' would now correctly Q/M-implicate disjoint refernce with the prior O object of the verb 'take'.

However, against that interpretation, there is the use of the same sentential frame 'O took them' as in 267 ut with different, indeed reversed, assignment of reference. In addition, the speaker provides gestural specification to the South, which does not accord well with the story so far: if this is an understanding check, it displays misunderstanding.

So what else could the speaker mean? Well, since in saying this he is gesturing towards the South, from where the policemen came, what he could be saying is that the policemen were returning to base. In that case, whom would they be taking? The culprits of course. So this interpretation is:

(66) Interpretation 2 for line 268
  (a) *07* took *them*2/3/4 back that way ((gestures South))
  (b) This clause is related to the prior ones by temporal sequence and causality and narrative finality

This interpretation requires that the zero-subject of a prior sentence becomes the zero-subject (or -agent, since this is an ergative language) of this sentence, but that just what one expects (the subject of 267 being I-linked to that of 268. A reason to think that this is exactly what the speaker R meant is that he has been competitively collaborating with B in the telling of this story, and this would be a *candidate story closing*, of the kind 'So the Mounties got their man'. If so, we would also have built an I-implicature connecting the two clauses, line 268 and its preceding one: they found the body *and then*, *as a result* incriminated the culprits and took them back South. Also in favour of this interpretation is that the gesture to the South would now be a sensible one, and GY speakers are very careful about their directional gestures — indeed they are so reliably oriented at all times that they scarcely ever make directional or gestural mistakes.

In the event, it is clear that B does not take the utterance on interpretation 2. The evidence for this is that in 269 he corrects the gestural specification

of motion to the South, by saying 'that way' and pointing clearly further to the West. Now in 267 B has already indicated by gesture that the whole party was heading Westward, after the police had come up from the South to collect the men who knew where the body was. He is now reasserting that motion, and in doing so, reminding us who can lead the police tothe body. In short, the single gesture specifying direction serves to insist that the zero-subject (or agent) of 'take back' is, jointly, the criminals, and the object (or patient) the policemen — i.e. B's contribution is built on interpretation 1 (as in (65)) of R's prior utterance. Note that in the following line R concurs with the Western motion and thus with this interpretation, apparently abandoning his bid for story closure (interpretation 2).

What is interesting about this 'ambiguity', which momentarily troubles the participants, is that it arises from the way in which the utterance in 268 is a truly equivocal object: in some ways it predisposes to one interpretation, in other ways to the other. But we can only see its equivocal nature by reconstructing, inter alia, the conflicting I- and Q/M-implicatures from the reference forms employed. Also of great interest is the demonstration of the theme earlier enlarged upon: *the inference of the reference of the valents of the verbs employed is deeply connected to the inference of inter-clausal linkage*. Although GY is no doubt exceptional in the minimal use of explicit valents and explicit clause-linkages, there is a general moral: there is no account of anaphoric linkage divorced from an account of inter-clausal linkage, and thus no isolated problem of anaphora. Finally, the example makes clear the rather special role that gesture plays in GY: it is hardly surprising that a language with zero anaphora and no verb agreement would find an ancillary channel of gestural information very useful indeed, displacement into that non-verbal channel itself counting as minimization.

I have dwelt at length on Guugu Yimidhirr because it neatly encapsulates my central theme: the impact of a principle of minimization in expression, with its interpretive corollary of maximization in inference, on both the structure and usage of language.

## 5. Conclusions

*Summary*

We started out from an anomaly in the Gricean programme, which led to the recognition of two competing principles governing the communication of information, and thus guiding interpretation, the I-principle and the Q-

principle. On the basis of intra-sentential data alone, it is hard to see quite how the clash between these two principles is systematically resolved, although we were able to formulate a resolution principle ((23)) that seems descriptively adequate. It amounts to the idea that the I-principle's expansion of informativeness is effectively bounded by the use of expressions that Q- or Q/M-implicate the complementary interpretation. This model of a number of competing pragmatic principles in systematic interaction is in marked contrast to other recent proposals, notably by Sperber & Wilson (1986), in favour of a single powerful pragmatic principle. There are definite advantages in being able to use one principle to bound the inferences generated by other principles, an option not open to the pragmatic monist.[24]

We then turned to the empirical tradition of CA to look for what evidence there was, in the sequential delivery of utterances, for principles of these sorts. It turns out that, naturally, the control of information is infinitely more complicated than our two principles might suggest; there is, as I have put it, a veritable social economy of information. Nevertheless, our two principles or their analogues do show through, and I concentrated on showing that the I-principle and its associated maxim of minimization might be discerned in operation in a whole host of conversation organizations, from the truncation of sequences to the specific inference mechanisms of 'membership categorization devices'. In addition, preference organization was shown to instantiate Horn's 'division of labour', the opposition between minimal forms and their stereotypical I-implicatures, and marked forms and the complementary interpretations. But the crucial discovery is the parallel to the Q- vs. I-conflict in the domain of reference to persons, for there is here a suggestion for another mode of resolution between conflicting principles. The solution is that resolution is *interactionally achieved*; that is, one can try a minimal form and see if it works, if not, escalate. We could formulate this in the slogan: "Try letting the I-principle win in the first instance, i.e. go for minimal forms; if that doesn't work escalate step by step towards a Q-principle solution".

In this review of relevant CA work, we concluded with findings about reference to third persons. This focus might have been held throughout, but it would have narrowed the scope of our review. However, reference to third persons is a crucial domain, and we then turned to see whether our principles had any application to the syntax of referential NPs. A quite surprising result seems to be that Binding Conditions B and C of the Government and Binding framework simply fall out of the systematic interaction of our three Neo-Gri-

cean principles. Further, some aspects of the Theory of Control, which is supposed to govern the interpretation of NP-gaps in ungoverned positions, also seem to be predicted by the three pragmatic principles; the 'Avoid Pronoun Principle' is likewise a natural consequence of our pragmatic apparatus. All this amounts to a potential pragmatic reduction of significant portions of Government and Binding Theory. But we should be cautious: given our adoption of 'pragmatic complementarism', there can be no presumption, from the demonstration of a possible pragmatic motivation for a syntactic pattern, of the lack of need for a complementary syntactic explanation. In many ways, especially in the area of control facts, the patterns consonant with our pragmatic principles appear *in English* to be thoroughly grammaticalized; but what we can expect is to find those same patterns in other languages as mere preferred usages with favoured but defeasible interpretations (see Comrie 1985, Levinson n.d.). If this expectation turns out to be correct, then there is at least one important implication for Chomskyan theory: the Binding and Control principles do not necessarily reflect abstract properties of some hypothetical innate language faculty; their universality may be due to the fact that they are natural consequences of rational principles of language usage.

Finally, we turned to Guugu Yimidhirr, and (from an English point of view) its improbable syntactic properties of zero-anaphora without any verb agreement or much cross-clause syntactic linkage. In this language, our three pragmatic principles and their interaction appear to play an important role in determining the reference of NPs and NP-gaps in discourse: given the syntactic freedom to omit NPs, such zero valents I-implicate coreference, while overt pronouns tend to Q/M-implicate disjoint reference. Where there are competing possible co-referential links, recourse is had to contextual likelihoods and to the systematic use of gesture, gesture itself counting (I have suggested) as a form of minimal further specification.

GY is the kind of language that the I-principle, if relatively uncontested and unconstrained, would predict, that is, a language where much has to be supplied by inference. The differences between GY and English in this regard seem very great, but it is important to remember that in conversation one does not find all the resources of English encoding automatically deployed. Indeed, quite to the contrary; as the CA work reviewed in 2. shows and as nicely exemplified by Silverstein (1985), minimization plays an important role in the use of English.[23] In any case, though, one is led naturally to sociolinguistic speculations about the kind of communities that support dif-

ferent kinds of language varying on these dimensions, speculations already entertained by Kay (1975) and others, who have related implicit communicaton systems (read 'I-languages') to small scale speech communities, and explicit communication systems (read 'Q-languages') to complex speech communities where participants cannot be presumed to share the same knowledge base. One is also reminded forcibly, by examples like (54)c. above from the legal register of English, that principles of minimization are very much matters of context, style and register. Our speculations have linked three domains: Gricean models of inference, conversation analysis and syntax; a natural fourth domain would be sociolinguistics, but that would be another paper.

## NOTES

1. *Acknowledgements*: This work derives in no small part from my exposure to an Australian language (see 4.); I am extremely grateful to John Haviland (who recommended the study of Guugu Yimidhirr anaphora to me and provided the grammatical background for it), Tulo Gordon, Roger Hart, the late Jack Bambi and all the people of Hopevale for making that possible, as well as the Australian Institute for Aboriginal Studies and The Australian National University for financial and logistical support. It also builds directly on collaborative work with Jay Atlas, without whom what there is in the way of theoretical apparatus underpinning this paper would scarcely exist. Larry Horn's generalization of those ideas, and indication of their possible application to GB theory, has also been an immediate stimulus. I owe a special thanks as well to John Heritage and Emanuel Schegloff for coming to immediate assistance during problems with the drafting of section 2. (whether they have rescued me from error only they can say). Later important comments from Jay Atlas, Phil Johnson-Laird and Tanya Reinhart I have scarcely been able to do justice to, since another paper would have resulted, but the existing paper has much benefited. My thanks too to my student Y. Huang, whose parallel work on Chinese zero anaphora made me realize that I would have to work fast if I was to say anything sensible about zero anaphora before he said it. Last but not least, thanks to Marcella Bertuccelli Papi and Jef Verschueren, whose organization of a conference on the curious theme of the integration of pragmatics made me think how these fragments might perhaps fit together after all.

2. Strictly, the speaker implicates that he knows that (2); see (12) below, and Gazdar (1979). Throughout, I shall be casual about the epistemic formulation, although it is important to formalization — see exposition in Levinson (1983: 135).

3. For difficulties concerning the definition of a Horn-scale see Harnish (1976: 362f, fn.46); Gazdar (1979: 58); Atlas & Levinson (1981: 44ff); Burton-Roberts (1984). The difficulty is that the entailment analysis is a necessary but by no means sufficient condition on scale-hood.

4. Incidentally, Horn (1985) lists indirect speech acts as further instances of the same kind of inference; Atlas & Levinson (1981: 37) while noting that indirect speech acts are ampliative inferences, refrained from claiming that they are I-implicatures (1981: 43) — they seem to require

essential reference to the purposive reasoning associated with Relevance; see 1.5 below. Atlas points out (personal communication) that the 'literal' interpretations of indirect speech acts seem if anything *more* informative than the corresponding 'indirect' interpretations, which would seem to rule out an account in terms of an informational maxim.

5. In Atlas & Levinson (1981), the phrase used was 'more precise'; the change in terminology does not indicate a change in conception.

6. By 'm-intended' I refer to Grice's concept (1957, 1968) of a 'meaning intention', where a speaker produces an utterance with the intention of inducing a specific belief in the addressee by means of the recognition of this intention.

7. A point underscored by the fact that synonymous, or near synonymous expressions (as in the examples in (21) and (22)), should give rise to identical implicatures *unless* the implicatures are derived from the maxim of Manner (I am grateful to Jay Atlas for reminding me of the pertinence of this point).

8. Tanya Reinhart (personal communication) finds this resolution schema too cumbersome to be plausible, suggesting that there must be some unitary principle that will subsume the Q- and I-principles, especially as both induce inferences more informative than what was said. Her suggestions deserve more attention than I can give them here. Meanwhile this Ptolemaic system should at least invite a Copernican solution.

9. Sperber & Wilson (1986) appeared as this paper went to press; at first blush, though, these comments seem just as germane to the latest version of their theory.

10. A further limitation of SW R is made explicit in Sperber & Wilson (1986: 36-37), where they indicate that their framework is primarily geared to *particularized* (context-induced) conversational implicatures. But it is the *generalized* (default) implicatures (which they dismiss as untypical) which are of central interest to linguistic theory, and which are the subject of this essay.

11. I am most grateful to Emanuel Schegloff for remarks that prompted the observations in this paragraph.

12. Emanuel Schegloff provides me with another kind of evidence. When a speaker has a 'word-search' problem, and cannot find a name for someone, he may use a description which allows the recipient to recognize the referent. Yet, despite the fact that recognition is now achieved, neither may be satisfied until the minimal reference form is found, indicating that recipient design (motivating the goal of referent-recognition) and minimization (motivating the search for a single minimal reference form) are independent, and independently satisfiable.

Incidentally, I should record here that my interpretation of Sacks & Schegloff (1979) was not exactly (Schegloff informs me) what was intended: by 'minimal reference form' the authors had in mind "a single reference form" (1979: 16), which might equally be a name or a description. However, the word-search kind of evidence (just discussed) for the two principles suggests strongly that the generalization of the notion 'minimal reference form' so that long descriptions, short descriptions and names are ranked as increasingly minimal is indeed correct.

13. Guugu Yimidhirr gesture plays a special and systematic role in reference; see section 4. below, Haviland (n.d.), Levinson (1984). Here a point of interest is that we have already suggested that there is ranking of channels as far as minimization is concerned, from gestural to linguistic, and, as developed in 4., zero anaphors in this language often get gestural reinforcement.

14. The need to demonstrate the I-principle derivation of pronominal coreference was made clear to me by remarks made by Ann Farmer and Robert Harnish on the Viareggio beach, and by

very detailed comments by Jay Atlas on a draft. Atlas also provided me with the Popperian argument and the argument at the level of sense, and a number of other improvements to this section. I am, as usual, greatly indebted to him.

15. Since (a) PRO would fall afoul of the 'equal lexicalization' constraint, (b) PRO is, arguably, a *gap* rather than an abstract item with meaning, and (c) PRO vs. pronoun do not form a scale of differential informativeness, since PRO in some locations needs not be co-referential with another NP.

16. Disregarding Chomsky's PRO as Specifier of NP and PRO in COMP (1981: 65), etc.

17. They also build in innumerable ways on Haviland's largely unpublished lexicography and discourse analysis of the language, which he has most generously shared with me. My work on the language is, compared to his, quite amateur and based on inadequate fieldwork, and we may look forward with great interest to the definitive treatment by him of many of the issues raised here.

18. Although deletion is free, the interpretation of the NP-gaps is naturally less so: there are analogues of English 'subject-control' and 'object-control' predicates, many of which patterns are documented in Haviland (1979: 135ff). But it is clear both from Haviland's discussion (p.137) and from other data that these principles of interpretation are only partially grammaticalized, requiring a pragmatic rather than a grammatical account. Such an account is provided quite simply by the framework of I-, Q- and Q/M-implicatures developed in this paper (Levinson n.d.).

19. I am most grateful to Roger Hart and the late Jack Bambi who told these stories for my benefit, and to Roger Hart and Tulo Gordon for helping me, with the patience of Job, to transcribe them. I am grateful too to John Haviland for demonstrating that film was essential and showing me how to use it, and for countless other kinds of assistance. Essential advice, introductions and assistance came too from Leslie Devereux, the Council of Hopevale, and above all many individual residents of Hopevale. Fieldwork was done with Penelope Brown, without whose collaborative effort the work could not have been accomplished. Copies of the films are in the Australian Institute for Aboriginal Studies, which body funded the fieldwork jointly with the Australian National University.

20. The pleonastic pronoun accompanying the noun in 172 (cf. also 170) (cf. also 193) is standard GY usage, a rather exceptional redundancy in this language that seems so much to favour the maxim of minimization! One way of looking at this usage is as parallel to the use of generic nouns (like *bama*, 'person', *mayi*, 'vegetable food') that often accompany specific nouns in GY much in the manner of classifiers, but which can occur alone as semantically general quasi-pro-NPs; the pronouns are then quasi-classifiers for animate entities (since they are restricted to that kind of referent). But perhaps the main function of these pleonastic pronouns, as Haviland has pointed out, is dependent on the fact that personal pronouns are inflected on a Nominative/Accusative basis (so that, e.g. a Nominative pronoun can be either a subject of a transitive or an intransitive verb), while nouns are inflected on an Ergative/Absolutive basis (so that, e.g., an Absolutive noun can be either a subject of an intransitive verb or the object of a transitive verb); this means that the conjunction of both case marking systems makes absolutely unambiguous which grammatical function the complex NP has (an object of a transitive verb, for example, being marked as both Absolutive and Accusative).

21. I have to confess to not understanding why the pronoun *nhangu* (Third-person, animate, Dative, but extended in recent usage to Accusative) is overt here. It's quite normal to have a zero agent and zero patient in transitive clauses, as in 197.

22. Incidentally, just prior to this extract there is a person reference sequence of just the sort we

mentioned in section 2.6, establishing recognition of Tracker Monday, which explains the use of the zero Np here.

23. One can find, for example, many uses of zero-subjects of main clauses in conversational English; see Ferguson, 1983 for such ellipsis in sports commentaries.

24. Sperber & Wilson (1986), and more directly Carston (1985), raise a puzzle for the kind of account I have given: how can we maintain a distinction between what is 'said' and what is 'implicated' if implicatures are used to fix reference (as in I-implicatures to coreference) and thus to determine what is said? (Grice's distinction is none too clear; see Harnish, 1976 for discussion.) A further anomaly follows too: if implicatures fix reference, and thus help to determine what is 'said', and what is 'said' may then have further implicatures (as in irony), implicatures look to be transitive — from which certain incoherences seem to follow (Burton-Roberts 1984; Atlas 1984: 359, 372-373).

The way out of this conceptual morass would seem to be to avoid defining implicatures in opposition to what is 'said' (certainly on a simple-minded definition, anyway), but rather by reference to the apparatus that induces them. We have long had the model of indexicals before our eyes, where pragmatic information plays a crucial role in specifying truth-conditions; and the role of implicatures in this process has long been recognized (see references in Levinson 1983: 34-35). For this and other reasons, most philosophers would insist that it is assertions (pragmatic entities) not sentences or semantic representations that are the bearers of truth-conditions. It would be possible to define two kinds of implicature, according to whether they are calculated on the basis of semantic representations (in the case of reference-fixing implicatures), or on the basis of truth-conditions (in the case of implicatures like irony). But the important observation is that the apparatus of maxims that predicts the two kinds of implicatures appears to be identical in both applications.

Appendix 1 Guugu Yimidhirr Narrative: Two Extracts

The texts are transcribed in the practical orthography described in Haviland (1979), with a partial morpheme gloss as follows. Lexical NPs are in the Absolutive (zero-realized) case unless marked otherwise, and pronouns are in the Nominative case unless marked otherwise. Other cases are marked as follows: ACC = accusative, ERG = ergative, LOC = locative, INSTR = instrumental, DAT = Dative; other glosses are: PART = particle, ASP = aspect marker, PAST = past tense marker, REFL = 'reflexive'/antipassive marker, PL = plural, REDUP indicates reduplicated morph.

Gestures in the right hand column have the approximate position of their 'acme' indicated by the symbol ( in the verbal transcript. Directional gestures are indicated by the customary abbreviations of the cardinal points; iconic gestures have the probable significance indicated in double quotes.

(63) From Revengl.arg:

| | | | GESTURES |
|---|---|---|---|
| 167 | B: | (dhanaagu miil nhaadhaadhi gurra.<br>dhana-:gu miil nhaa-dhaadhi gurra<br>3PL+NOM-EMPH eye-ABS see-REF+PAST also<br>They looked at each other | (2 hands up "look at each other"? |
| 168 | | garrbaadhi gurra.<br>garrba-adhi gurra<br>hold-REF+PAST and<br>Then they grabbed each other | |
| 169 | | (warra nyulu yii- (you know<br>warra nyulu yii you know<br>old 3sNOM this<br>Old - you know | (W (thumb W |
| 170 | | bama nyulu nhayun warra yabarraban<br>bama nyulu nhayun warra yabarraban<br>man 3sNOM that+ABS ~ giant<br>That chap was really gigantic | |
| 171 | R: | yabarraban.gu<br>yabarraban-gu<br>giant-EMPH<br>A real giant. | |
| 172 | | bama nyulu dyimili-mi gaarga<br>bama nyulu dyimili-mi gaarga<br>man 3sNOM Jimmy Lee~DAT younger brother<br>He was Jimmy Lee's younger brother | |

173 B: yeah i.) (maani yindu ('pick up & dump'
yeah maani yindu
get+PAST other+ABS
Yeah. He grabbed one of them

174 dalmba-gabay ((1.0) ('box'
dalmba-gab-ay
wrestle-zap-PAST
and threw him

175 R: yindu dalmba-gabay
yindu dalmba-gab-ay
other+ABS wrestle-zap-PAST
Then he threw one of them!

176 B: (yindu dalmba-gabay well b- ('box'
yindu dalmba-gab-ay
other+ABS wrestle-zap-PAST
Right! Then he threw another

177 (bulii-bulii galmba you know (points down N
bulii-bulii galmba you know
also
Bulii-Bulii was also you know,

178 (ngaanaarru (warra= ('empty hands'; (shoulder jogs 'limping'
ngaanaarru warra
whatchamacallit very
what do you call it?

179 R: =helpless
helpless eh?

180 B: (helpless nhaadhi ('limping'
helpless nhaa-dhi
see-PAST
Yeah, helpless

181 R: aa
Yeah
[

182 B: (nhangu maani ('pick up & dump'
nhangu maani
3sACC get+PAST
He (Douglas) took him (Bulii-Bulii)

183 nhangu galmba dalmba-gabay
nhangu galmba dalmba-gab-ay
3sACC also wrestle-zap-PAST
and threw him too

184 by the time nhangu (nhayun gundaarnday (points down N
by the time nhangu nhayun gund-aarnd-ay
3sACC that+ABS hit-REDUP-PAST
While he (Douglas) was hitting that third chap

185 nyulu nhila (yii baaa dyibaarr (points down S
nyulu nhila yii baaa dyibaarr
3sNOM now this man South+ALL
He (Bulii-Bulii) this chap to the South

186 dabaarr-manaadhi
dabaarr-manaadhi
good-become+PAST
came to

187 R: (yandayqu? (nods S
yand-ay-qu
arise-PAST-EMPH
Did he get up?

188 B: yanday (marrgin nhanaay ("fumble around & find"
yanda-y marrgin nhanaa-y
rise-PAST gun+ABS take-PAST
He got up and grabbed a gun

189 (marrginda gunday ("hold gun"
marrgin-da gunda-y
gun-INSTR hit-PAST
He (Bulii-Bulii) shot him (Douglas) with the gun

190 R: iii
wow!
Ahah!
[
191 B: ((oof) (.) (panaa budhu ("collapse"; ("collapse"
panaa budhu
Bang! alright intensifier
Bang! Okay then

192 R: badhaadhi
badha-adhi
finish-PAST
That finished him

193 B: (yirngaadhi gurra ("gather around"
yirnga-adhi gurra
encircle-PAST and
They stood around him (Douglas)

bulii-bulii-ngun (naabal maani (points N
bulii-bulii-ngun naabal maani
Bulii-bulii-ERG stone get+PAST
Bulii-bulii took a stone

194 R: (dumu baydyarrin (B gestures 'throw down'
dumu baydyarr-in
chest+ABS crush-PAST
He crushed his chest

195 B: ((Bangs chest)) ('crushes chest'

196 ((2.5) (empty hands 'finished'

197 B: (miidaarrin ('lifting'?
miidaarr-in
lift-PAST
They lifted him

198 maandi (guwaalu (arm slowly raised to point W
maandi guwaalu
take+PAST West-LOC
and carried him away to the West

199 (nhayun gala (guwagu birrii nhayun maynggu wunaarna bada (W; (W edge of river bank
nhayun gala guwa-gu birrii nhayun maynggu wunaa-rn-a bada
that+ABS far West-LOC-EMPH river-ABS that+ABS mango tree lie-REDUP-PRES down
far West to the river where the mango trees are

(64) Extract 2: Tracker Monday and the police find the murderers and their victim

266 B: ((1.0) gaday (gurra) (E, E
gad-ay gurra
*come-PAST also*
So (Tracker Monday) came too (with the police)

267 maandi dhanaan ((2.0) (maandi (W; (W
maand-i dhana-an maand-i
take-PAST 3PL-ACC take-PAST
and (they) took them (the culprits), took them (and said)

(miirriila wanhdhaalbi (duugay yurra (fingers down "you guys"; (W
miirriil-a wanhdhaal-bi duug-ay yurra
show-IMP where-LOC dig-PAST 2PL+NOM
"Show us where you buried him"

268 R: dhanaan back (yarrba maandi? (points over shoulder S
dhana-an yarrba maand-i
3PL-ACC that way take-PAST
So (they) took them back that way

269 B: (yarrba (from E to W
yarrba
that way
No, (the police took the criminals) that way

270 R: aa yarrba guwa
aa yarrba guwa
ah! that way West
*Ah, that way, West, I see*

271 B: aa
aa
yes
Yeah

272 R: (bubu dhumbuurgu **ɨiirriilin** (yii bada (W; *(points down to ground*
bubu dhumbuurgu **ɨiirrii**-in yii bada
land-ABS straight show-PAST here down
(The criminals) pointed straight down to the ground "Here!"