# *Game theory and communication*[*]

NICHOLAS ALLOTT

## Abstract

This paper assesses recent game-theoretic accounts of communication which are motivated by the goal of understanding links between rationality and communication, particularly Prashant Parikh's model (Parikh 1991, 2000, 2001). Following an account of the model, it is argued that it faces empirical and theoretical problems. Comparisons are made with relevance theory (Sperber & Wilson, 1986/95) and improvements to the model are suggested. An alternative kind of approach, represented by the work of Lewis (1969) and van Rooy (forthcoming) is briefly discussed, as are doubts about the formalisation of rationality offered by game theory.

## 1 Introduction

This paper examines a serious attempt at a post-Gricean game theoretic account of communication. Doubts will be raised about some of the aspects of the proposed treatment and the suggestion will be made that at least some of the problems that the proposal faces are problems that stem from using game theory in this field. I argue that other problems may be due to the adoption by the theory of views about language and communication which are common but arguably mistaken. Much of this paper is rather negative in tone, therefore, but the intent is to be constructive in criticism. I do not want to suggest that a game theoretic treatment of communication cannot or should not be attempted: rather, the aim is to raise serious empirical and theoretical questions which I think need to be addressed for that research programme to make progress.

[*] The author wishes to thank Deirdre Wilson for her unstinting support including comments on successive drafts of this paper and discussions of several connected topics. Thank you also to Richard Breheny, students at RCEAL and participants at ESPP 2003 for discussion. The mistakes are my own.

## 2 Why employ game theory in an account of communication?

Game theory is concerned with strategic interactions, that is, situations where two or more players (or agents) have to make decisions and the way things turn out for each player may depend on the other player or players' choices.

Superficially, at least, human communication looks like this sort of situation[1], since a speaker makes an utterance and a hearer tries[2] to interpret it. We say that the hearer *tries* to interpret the utterance because a great deal can be meant by the speaker which is not encoded in the linguistic form (the lexical items and the syntactic structure) of the phrase or sentence uttered. The hearer must, at least, choose a meaning for ambiguous expressions, assign reference to indexical elements such as pronouns, decide on reference class and scope for quantifiers, work out for some lexical items how loosely or narrowly they are being used and recover intended implicit meaning ('implicatures').

How successful the speaker and hearer each are seems to depend on choices made by the other: if the interpretation the speaker intended is (close enough to) the one the hearer works out, then communication is successful, otherwise there is miscommunication. This apparent degree of shared interest has led to the suggestion that communication be modelled as a coordination game (Lewis, 1969; Parikh, 1991). (See below for coordination games)

### 2.1 Game theory, communication and rationality

A central attraction of game theoretic models of communication is that they might help to elucidate the link between rationality and human communication (if there is one). Grice was convinced that principles governing communication would fall out from general assumptions about rationality and the cooperative nature of communication (perhaps when taken together with other assumptions about human beings or the communicative situation), but he was unable to show this.

> I am… enough of a rationalist to want to find a basis that underlies these facts [about people behaving as the conversational principle and maxims say]… . I would like to think of the standard type of conversational practice not merely as something that all or most do IN FACT follow but

[1] See for example Parikh (1991, p. 473) and Sally (2003, p. 1223) for statements of this intuition.

[2] Use of the word 'tries' is not meant to imply that the processes involved need be conscious or available to introspection, either during processing or subsequently. The same caveat, standard in cognitive science, applies to other verbs I have used in describing the hearer's task, including 'choose', 'decide', 'work out' and others.

> as something that it is REASONABLE for us to follow, that we SHOULD NOT abandon. (Grice, 1967, p. 48, his emphases.)

Game theory has certain assumptions about rationality built in, and a game theoretic account of communication would inherit these assumptions. This approach promises, therefore, to make the link between rationality and communication clearer and more precise. On the other hand, a game theoretic model of communication would also inherit any empirical and theoretical disadvantages of the particular formalisation of rationality adopted by game theory. (See section 7.2 below for further comments.)

There are at least two ways of making the link between game theory and communication (van Rooy, forthcoming, section 1): either rationality considerations are taken to apply to languages or directly to the communicative situation. Lewis' work and recent work by van Rooy is in the former tradition; Parikh's model (Parikh, 1991, 2000, 2001) takes the second approach, looking at how communication is achieved between a speaker and a hearer with the language as a given. Below, I examine Parikh's model, arguably the most developed attempt at either of the approaches, and in my opinion the most promising attempt at a game theoretic treatment of communication. I will compare this model with the account of communication given by relevance theory, with the aim of assessing how well Parikh's account compares with modern cognitive pragmatics in its explanation of retrieval of explicit and implicit meaning.

## 2.2 Criticisms of Parikh's model

The criticisms I present in this paper fall into four categories. First, I will examine cases where Parikh's game-theoretic approach seems to get the data wrong. Secondly, I will argue that it makes appeal to intuitions of relevance without fully acknowledging this. The point of this criticism is not that an account of communication should avoid use of a concept of relevance: rather that a formal model should make clear its dependence on such a key concept, instead of using it implicitly, partly so that fair assessments can be made of the relative complexity of different theories, but mainly to avoid assuming solutions to problems that a theory of communication should be centrally concerned with.

My third criticism is that Parikh's account generates huge structures—even for communication involving inferences about the explicit meaning conveyed by fairly simple sentences—and thus faces one of two problems: either it leads to a combinatorial explosion if the model is taken as a model of mental representation or processes, or, if the model is taken instead to describe the logic of the communicative situation, it makes comprehension look like a very tough problem

for the (unspecified) mental mechanisms which have to arrive at the same solutions as Parikh's model but in real time with much less apparatus.

Finally, I want to suggest that Parikh's account may be the wrong kind of account of communication insofar as it has certain theoretical commitments which are also often found in philosophical accounts of communication or meaning and in recent work on formalising pragmatics. For example, Parikh assumes that language is completely expressive, so that for any thought which can be communicated using language there is at least one sentence which unambiguously expresses that meaning. This is closely related to Katz's thesis that thought is effable (Katz, 1981). It is arguably implausible at the level of individual concepts (Sperber & Wilson, 1997/98) as well as at the level of sentences and propositions (Sperber & Wilson, 1986/95, pp. 191-3; Carston, 2002, pp. 32ff.). Perhaps more seriously, this view of the relation between language and thought is connected with theories of communication which see it as simply a way of transferring a proposition from the mind of the hearer to the mind of the hearer. Sperber and Wilson have argued that theories of this type underestimate the importance of the role of inference in human communication (Sperber & Wilson, 1986/95, pp. 1-15 & 24-28).

I think that the problems the model faces are connected. I will mention some but not all of these connections in discussing the problems.

## 2.3 Competitive and cooperative games

Scissors, paper, stone is a competitive game with two players, each of whom has three choices at each turn. The possible outcomes are *win* (W), *lose* (L) and *draw* (D).

| | | player 2 | | |
|---|---|---|---|---|
| | | **scissors** | **paper** | **stone** |
| player 1 | **scissors** | D, D | W, L | L, W |
| | **paper** | L, W | D, D | W, L |
| | **stone** | W, L | L, W | D, D |

Figure 1 *Scissors, paper, stone*

In game theory, the outcome obtained by a combination of decisions is called a payoff. It is usual to give numbers to payoffs: positive for good results, negative for bad ones (and zero for those which are neither).

player 2

| player 1 | | **scissors** | **paper** | **stone** |
|---|---|---|---|---|
| | **scissors** | 0, 0 | 1, -1 | -1, 1 |
| | **paper** | -1, 1 | 0, 0 | 1, -1 |
| | **stone** | 1, -1 | -1, 1 | 0, 0 |

Figure 2 *Scissors, paper stone* with numerical payoffs

Scissors, paper, stone is a competitive game—each player does best when the other player does worst. Other situations are cooperative games, where the interests of the players are lined up.

Consider Figure 3:

player 2

| player 1 | | **Trafalgar Square** | **UCL front quad** | **…** |
|---|---|---|---|---|
| | **Trafalgar Square** | 20, 20 | -10, -10 | **…** |
| | **UCL front quad** | -20, -10 | 10, 20 | **…** |
| | **…** | **…** | **…** | **…** |

Figure 3 Meeting game

Here both players want to meet, so the outcomes where they make the same choice have high payoffs for both. The outcomes where they do not meet have negative payoffs to reflect the effort involved in making the journey. Also, player 1 has an aversion to UCL, so for him the payoffs at UCL are lower than elsewhere. If this were not the case, this would be a purely cooperative game. As we will see, Parikh models communication as either cooperative or nearly so.

In the simple game where Trafalgar Square (t) and UCL front quad (u) are the only choices, <t, t> and <u, u> are both 'Nash equilibria': situations where neither player will be better off if he changes his choice unilaterally. (I return to the concept of Nash equilibria in section 3.1.)

Games can also be played sequentially: first one player takes a turn, then the other, *with the second player knowing what the first player has chosen*. In this case, the games can be represented as trees. Figure 4 is a tree for a meeting game with two players and two choices each.

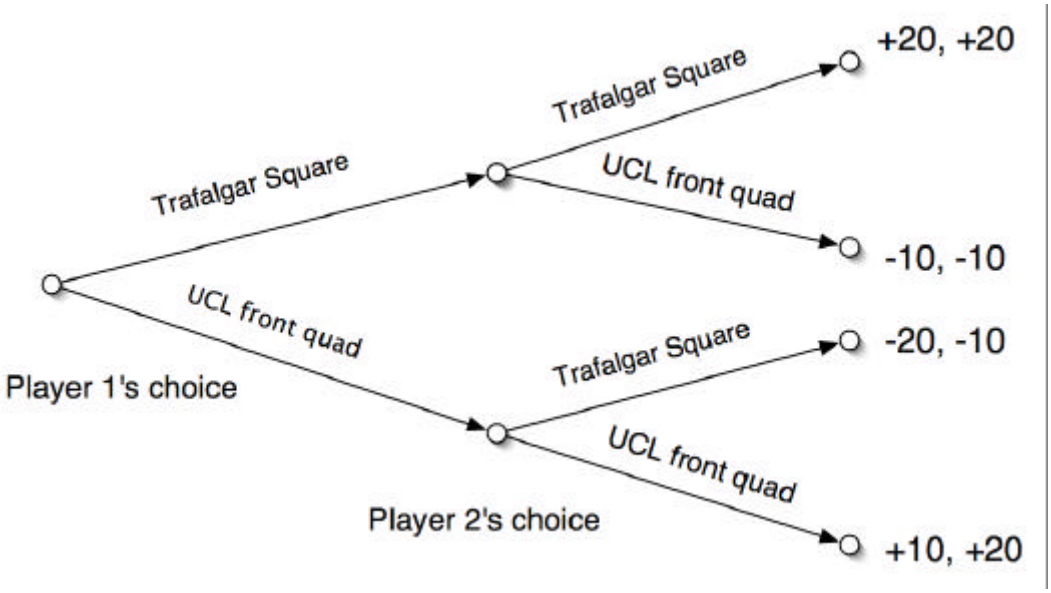


Figure 4 Sequential meeting game with two choices and two players

Here player 1 gets to move first. How should he choose his move? He looks at what player 2 will do in each sub-tree, that is, in each situation that player 2 could be in. Here it is simple: we assume that player 2 chooses so as to maximise his payoff. It follows that if player 1 chooses t then player 2 will choose t; if player 1 chooses u then player 2 will choose u (maximising his pay-off to +20 versus –10 in each case). The other outcomes can be ruled out. So player 1's choice is between <t,t> (the top branch), and <u,u> (the bottom branch). He prefers the pay-off of +20 for <t,t> to the +10 he would get for <u,u> so he will choose t and player 2 will then choose t.

The games we will be looking at are also sequential games: a speaker chooses an utterance, and a hearer, knowing what has been uttered but not the intended interpretation, chooses an interpretation.

## 3 Parikh's model

Consider a situation where a speaker makes an utterance and a hearer tries to interpret it. Parikh starts by examining cases where the sentence uttered has two possible meanings. This could be due to lexical or structural ambiguity, to the need to assign reference or to the purely pragmatic availability of two readings for the sentence.[3] Parikh uses the situation in which the speaker utters (1):

[3] Parikh regards all of these cases as cases of ambiguity, not restricting the term, as is more usual, to cases where two or more sets of linguistic items or linguistic structures correspond to the

(1) φ: Every ten minutes a man gets mugged in New York

According to Parikh, this has the two possible meanings:

(2) p: Every ten minutes someone or other gets mugged in New York.[4]
(3) p': Some particular man gets mugged every ten minutes in New York.

As Parikh acknowledges, there are other parts of the meaning of (1) which the hearer must resolve: 'New York' might mean the state or the city, and 'ten minutes' is clearly being used somewhat loosely. The aim is to show first how the model of communication resolves one uncertainty at a time, then allow for its extension to cover realistic cases where several aspects of the meaning must be fixed simultaneously.

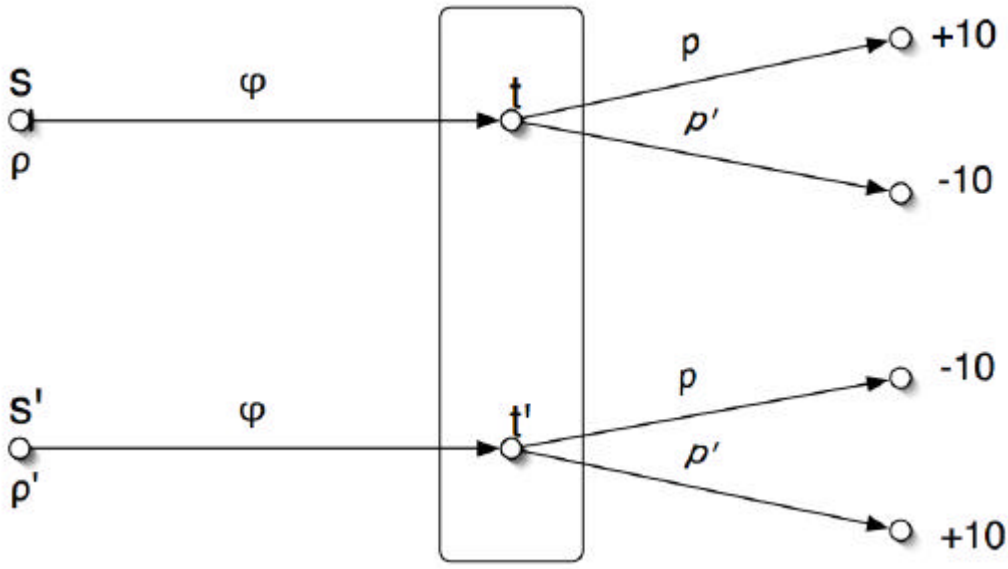


Figure 5 Part of local game for utterance of φ (Parikh, 2001, p. 30)

---

same phonetic form. I will use the term 'ambiguous' only in its narrow sense, since *disambiguation* marks a theoretically important category which is in contrast with reference assignment and pragmatic enrichment at least. Note however that it makes no difference to Parikh's account whether or not (1) is structurally ambiguous, corresponding to two representations which differ in quantifier scope, for example, given that it has the two readings given in (2) and (3).

[4] Although Parikh presents this sentence as unambiguous, it has the same two readings as (1). It may be problematic for Parikh that it is often difficult to find unambiguous alternatives, given his claim that successful communication depends on consideration of unambiguous alternative utterances.

Consider figure (5). There are two initial situations, s and s'. s is the situation in which the speaker means to convey p; s' is the situation in which she means to convey p'. In either case, she utters φ. After the utterance there are also two situations, t and t', where t is the situation where the speaker means to convey p and has uttered φ and t' is the situation where she means to convey p' and has uttered φ. The speaker, knowing what she wants to convey, knows which situation she is in. The hearer does not. His uncertainty is represented by the box around t and t'. He assesses the probability of s as ρ and the probability of s' as ρ'. He then chooses an interpretation—either p or p'. Note that his preferred choice depends on information he does not have. If he is in t he prefers to play p; if he is in t', he prefers p'. This is reflected in the payoffs. For the moment, Parikh assumes that successful communication is worth +10 to each player and unsuccessful communication is worth –10. This is based on two assumptions, both of which he later relaxes: first, that all information has the same value; secondly that the information is worth the same to both players.

According to Parikh, successful communication depends on consideration of non-ambiguous alternative utterances. In figure (6) the alternative utterances μ and μ' have been included. μ unambiguously means p and μ' unambiguously means p'. Therefore if the speaker utters μ, the hearer knows he is in situation e and will choose interpretation p. Similarly an utterance of μ' leads the hearer to choose p'. The payoffs are worked out by assigning +10 as before for successful communication, and -3 for the extra effort involved in the production and comprehension of these longer and linguistically more complex utterances. Parikh does not give details of the way the linguistic complexity translates into effort; later we will see that he allows the cost of constructing or processing a mental representation to come in also as a negative factor in payoffs.

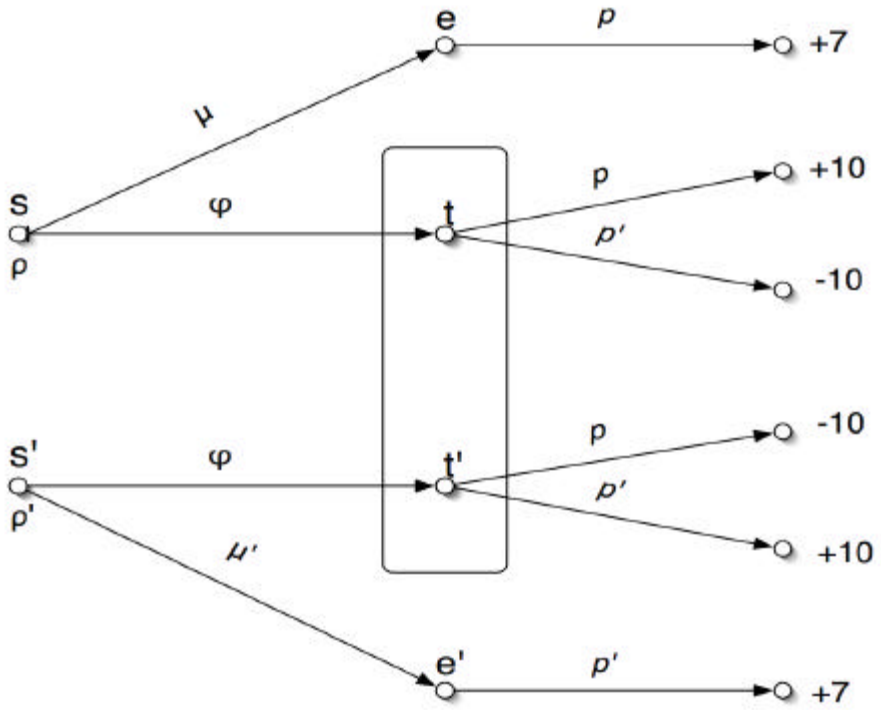


Figure 6 Local game with alternative utterances (Parikh, 2001, p. 31)

Figure 6, then, is the hearer's model of the interaction, which he constructs when φ is uttered. The speaker can also construct this model, since it is based on shared knowledge. If there is a unique solution to the game, therefore, it will be known to both players. Successful communication using φ will be possible. I return to the solution of the local game in section 3.1 below.

If the speaker knows the payoff of the local game she knows the payoff from uttering φ. She compares this with uttering μ and other alternatives as shown in the 'global game', Figure 7.

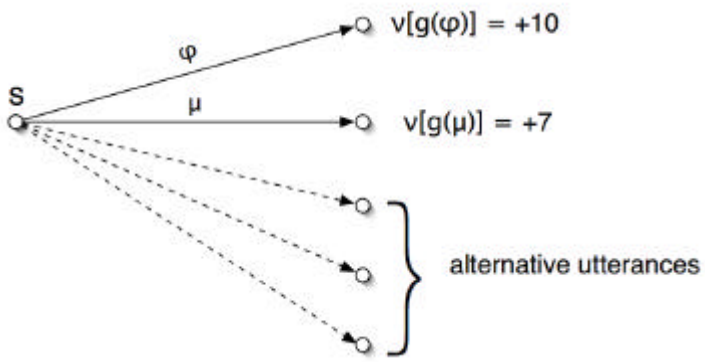


Figure 7 The speaker's choice of utterances (after Parikh, 2001, p. 32)

As shown, φ has a higher payoff than μ, so the speaker should choose it as long as it also has a higher payoff than alternative utterances.

## 3.1 Solutions for the game

Returning to the problem of solutions for the local game, consider the different possible strategies for both players. A strategy is a specification of the choices a player will make at all decision nodes. Here, the non-trivial choices are for the speaker at s and at s' and for the hearer at t or t' (the hearer's choice is constrained to be the same at t and t', since he cannot know which of them he is at). This gives eight different strategies, two of which are Nash equilibria, that is, solutions where neither player can do better by changing his or her choice unilaterally.

| **What A will do in s** | **What A will do in s'** | **What B will do in (t,t')** | **Nash equilibrium?** | **If not Nash equilibrium, why not?** |
|---|---|---|---|---|
| φ | φ | p | no | A should defect to μ' in s' |
| φ | μ' | p | yes | - |
| φ | φ | p' | no | A should defect to μ' in s' |
| φ | μ' | p' | no | A should defect to μ in s; B should defect to p |
| μ | φ | p | no | A should defect to μ' in s'; B should defect to p' |
| μ | μ' | p | no | A should defect to φ in s |
| μ | φ | p' | yes | - |
| μ | μ' | p' | no | A should defect to φ' in s' |

The first Nash equilibrium (N1) is the intuitively correct solution: if the speaker wants to mean p—the more likely meaning—then she says φ—the shorter but ambiguous utterance; if she wants to mean p'—the less likely meaning—then she says μ'—the longer but unambiguous utterance; and the hearer correctly interprets the ambiguous utterance φ as having the more likely meaning, p.

The second Nash equilibrium (N2) is a kind of inverse of the intuitively correct solution: if A wants to mean p—the more likely meaning—then she says μ—the longer but unambiguous utterance; if she wants to mean p'—the less likely meaning—then she says φ—the shorter but ambiguous utterance; and B correctly interprets the ambiguous utterance φ as having the less likely meaning, p.

It is clearly good that the other strategies are ruled out, since they describe even stranger arrangements than the second Nash equilibrium, but it is also necessary for Parikh to rule out this equilibrium, leaving only the intuitively correct solution. To do this he brings in another solution concept, Pareto-dominance. If some solution A

is better for at least one player than solution B and A is not worse than B for any player then solution A Pareto-dominates solution B. Pareto-dominance can pick out different solutions from Nash equilibrium. To avoid this, Parikh applies it as a secondary criterion to choose between Nash equilibria.[5] In this case, this works out as choosing the Nash equilibrium with the highest expected payoff. Using expected utility as the measure of the worth of an outcome, as is usual in game theory, the expected payoff for an outcome = payoff x probability of outcome. (Thus, for example, a rational agent should prefer a 50% chance of €100 to €20 for sure.)

Thus, setting $\rho = 0.9$ and $p' = 0.1$, we have the following payoffs for the Nash equilibria:

payoff of 'correct' solution = $10 \times \rho + 7 \times \rho' = 9.7$
payoff of 'incorrect' solution = $7 \times \rho + 10 \times \rho' = 7.3$

The intuitively correct solution Pareto-dominates the other solution, which can therefore be eliminated. Generally, N1 Pareto-dominates N2 iff

$$\upsilon(\varphi \text{ in } s, p).\rho + \upsilon(\mu' \text{ in } s', p').\rho' > \upsilon(\mu \text{ in } s, p).\rho + \upsilon(\varphi \text{ in } s', p').\rho'$$

where $\upsilon$(strategy) is the payoff for that strategy.

This is a criterion for successful communication, therefore.

---

[5] Van Rooy has made some game-theory-internal criticisms of this solution concept (van Rooy, forthcoming, section 4), saying that reaching Pareto-dominant solutions seems to require coordination between the players of the sort that can be achieved by a round of 'cheap talk' (communication with no costs) before the game. Naturally, cheap talk is not an option for Parikh on pain of an infinite regress, since what is to be explained is communication. It is true that Pareto-dominant Nash equilibria are not necessarily the ones that are chosen: in a coordination game with many Nash equilibria, all of which have equal payoffs except for one which is Pareto-dominated by all of the others, that one is a focal point and is likely to be chosen, despite its lower payoff. An example is meeting at UCL in the simultaneous London meeting game in the text.

Another problem for Pareto-Nash equilibrium as a solution concept might arise from recent research (mentioned by Sally, 2003, pp. 1229 f.) showing that players often prefer 'risk-dominant' (actually risk-avoiding) strategies to Pareto-dominant strategies. Still, I am not sure how serious these problems are for Parikh. It seems to me that he could say that the cognitive communication heuristics that find solutions just do look for Pareto-dominant Nash-equilibria. This would be an empirical claim.

## 4 Where do the probabilities come from?

The subjective probabilities assigned by the hearer to the situations s and s' are crucial in determining whether successful communication occurs in the model. A number of questions might be asked about these probabilities, particularly, where they come from, that is, how do the hearer and the speaker arrive at them? (Note that the speaker must assign at least similar probabilities to the ones assessed by the hearer for communication to be successful.)

Parikh's answer is that the probability of the speaker's wanting to convey p is related to the probability that p is true, although with a complication:

> Since it is common knowledge that p is more likely than p', we can take it as common knowledge that A [the speaker] probably intends to convey p rather than p'... In general, there is a difference between these two probabilistic situations, and it is only the second, involving A's intention, that matters. In the absence of any *special* information, one situation does inform the other... (Parikh, 2000, p 197, my emphasis.)

Two important questions are raised by this formulation. First, what is 'special' information and how do we know when it should override the probability of the proposition? Secondly, if the probability of the proposition is sometimes involved, why can Parikh not use that probability in all cases?

The answer to the first question is given in a parallel quotation from an earlier paper: "Note that in general, there is a big difference between the likelihood of a proposition's being true and the likelihood of an agent intending to communicate that proposition. *It is the absence of further relevant information that justifies identifying the two possibilities***.**" (Parikh, 1991, p 482, my emphasis.)

No characterization of relevance is given, so it appears that an intuitive notion of relevance is playing a crucial role in overriding the normal derivation of the probabilities. To summarise what Parikh is proposing here: in the normal case the probabilities ρ and ρ' come directly from the probabilities of p and p'. Note that there must be some kind of transformation of these probabilities since we assume the speaker is in either s or s', so $\rho + \rho' = 1$. The probabilities of p and p' do not in general add up to unity, since p and p' need not be mutually exclusive. For example, one of the propositions may entail the other, as in Parikh's example where (2) p: Every ten minutes someone or other gets mugged in New York. is entailed by (3) p': Some particular man gets mugged every ten minutes in New York. This transformation is not specified, but simple formulations might be tried, for example keeping the ratio between the probabilities the same but normalizing them so that $\rho + \rho' = 1$. In 'special' cases, however, according to Parikh, the probabilities derived

this way will not be correct and we modify them somehow to take account of information which is (intuitively) relevant.

Clearly this formulation is not very satisfactory since three key points are left unspecified: the transformation of the probabilities, the 'special' circumstances under which relevant information overrides these probabilities and the way in which this relevant information (perhaps in conjunction with the original probabilities) determines ρ and ρ' in these cases. Given these difficulties, one might ask whether Parikh needs to go beyond supposing that the correct probabilities come from the probabilities of the propositions. That is, does he need to suppose that there will be 'special' cases? In the next section I draw on work by Wilson and Matsui to show that he has to make this assumption *if* he is committed to using the probabilities of the propositions (normalized) as the probabilities of the situations s and s'. I also propose alternative strategies.

### 4.1 Truth, relevance and disambiguation

Wilson and Matsui examine accounts of pragmatic processes such as disambiguation which use rules such as 'The correct interpretation is the one which is most likely to be true.' Accounts of this type are called truth-based. These accounts often make incorrect predictions: in general the correct interpretation need not be the one most likely to be true, as examples such as (4) show:

(4) John wrote a letter.
    (a) John wrote a letter of the alphabet
    (b) John wrote a letter of correspondence
    (Wilson & Matsui, 1998, p. 24)

The lexical item 'letter' is ambiguous, so that (4) could mean either 4(a) or 4(b). The disambiguation naturally arrived at (at least in non-biasing contexts) is 4(b), but this cannot be more likely to be true than 4(a) since 4(a) is entailed by 4(b): anyone writing a letter of correspondence must be writing letters of the alphabet[6]

It is examples of this kind which rule out just using the probabilities of p and p' for ρ and ρ' in Parikh's model. Is it a good move to say that *normally* a truth-based approach is followed but that it needs to be modified in certain cases? Symmetrical examples like (5) and (6) suggest that it is not.

(5) Mary is frightened of <u>dogs</u>. (ambiguous between male dogs and dogs in general)

[6] Abstracting away from non-alphabetic writing systems, of course.

(6) Mary is frightened of cats. (ambiguous between cats in general (lions, domestic cats, tigers etc.) and domestic cats)
(Wilson, lecture notes. See also Sperber & Wilson, 1986/95, p. 168)

In (5) the intuitively correct reading has the more general sense of the term, *dog in general*, and is therefore the reading which is more likely to be true, so Parikh's treatment would work here without any appeal to special circumstances. In (6), on the other hand, the intuitively correct reading has the more specific sense of the ambiguous term, *domestic cat*, and this is therefore the reading less likely to be true. In this case, therefore, Parikh would apparently want to say that relevant information somehow corrects the probabilities of the propositions so that the probability that the speaker wants to convey *domestic cat* is higher than the probability that she means to convey *cat in general.*

There is a better solution available, however. In both cases, the intuitively correct meaning is the one with the more accessible of the two readings of the terms. *Accessibility* is a psycholinguistic concept. Empirical work in psycholinguistics aims to determine what makes a lexical item or a sense of a lexical item more accessible on one occasion or another. Well corroborated current theories propose that the frequency of use of the sense and the recency of its use are the key factors in accessibility of different senses in disambiguation. In a 'neutral' context there are no recent uses to consider, so accessibility for a reader encountering (5) and (6) as they are presented here would depend only on the frequency of the senses. Relying only on this cue for disambiguation would get both examples right: the more common meanings are *domestic cat* and *dog in general.*[7]

These kinds of examples are not particularly hard to find. Parikh's own examples raise the same questions. Recall that the explanation given of why we arrive at reading (2) p: Every ten minutes someone or other gets mugged in New York., for an utterance of (1) φ: Every ten minutes a man gets mugged in New York, is that the alternative reading—in which it is a particular man who gets mugged—is less likely to be true.

According to Parikh, the ambiguity in (7) is ineliminable (in some contexts only, I assume):

---

[7] As Wilson (p.c.) points out, "Sperber & Wilson's notion of relevance sheds some light on why those particular senses of 'letter', 'dog' and 'cat' are the most accessible ones. The fact that someone wrote a letter in the 'letter of the alphabet' sense would very rarely be relevant enough (achieve enough effects, at low enough effort) to be worth mentioning, so 'wrote a letter' will rarely be used in that sense, and this explains its infrequency of use. Similarly, narrowing 'dog' to 'male dog' would rarely make enough difference to relevance to be worthwhile. By contrast, 'cat' in the general sense covers such very different sub-cases that it may make a huge difference to relevance to know which sub-case is involved. So 'accessibility' isn't a magic potion but something that is (a) empirically testable and (b) often theoretically predictable."

(7) A comet appears every ten years.

(Parikh, 2001, p. 41)

The claim is that here $\rho = \rho' = 0.5$, so the expected payoffs for N1 and N2 are equal and the model predicts that the utterance should be unresolvably ambiguous. This is slightly problematic, since as before, the more general meaning ('Some comet or other appears…) is entailed by and therefore at least as likely to be true as the other ('A particular comet appears…'). Therefore the model *can* set the probabilities equal without overriding the logical relationship, but this is only one end of the allowed probability ratio and it is not clear how the model is supposed to arrive at it.

In examples where the more specific reading is the intuitively correct one, such as (8), the model will have to appeal to special information to predict this, since the same entailment between the propositions expressed by the general and the specific reading holds and the specific reading cannot be more likely to be true than the general one.

(8) A comet appears every 76 years.

Perhaps a better move for a supporter of Parikh's model would be to say that the probabilities $\rho$ and $\rho$' reflect the psycholinguistic activation and accessibility data rather than the probabilities of the propositions the speaker wants to convey. This would allow the model to make correct predictions at least in cases like (4),(5) and (6) without appeal to special information.

Example (9), however, shows that this move on its own might not be the best way for the model to make correct predictions of explicit meaning generally.

(9) Johnny is a boy who is with his friend Billy. He calls home to ask if he can go swimming with Billy. His father answers the phone. It is ok with the father if Johnny goes along but he wants to check with his wife, so the father calls out to Johnny's mother:

> Johnny's father: Johnny wants to go swimming with Billy.
> Johnny's mother: He has a cold.
>
> (Breheny, p.c., from Breheny, 1999)

As Breheny says (p.c.), "The intuition is that the mother's reply is understood to be referring to Johnny and this would also be true if the father instead calls out, 'Billy wants to take Johnny swimming'". This is important because according to the best current theory of accessibility for reference assignment, centring theory, the subject of a previous sentence should be more accessible than the object. Varying the

position of 'Billy' and 'Johnny' in the utterance before the pronoun controls for at least some accessibility factors, therefore. It seems, then, that getting the disambiguation right here is not simply a matter of accessibility, at least as it is in centring theory.

An advocate of Parikh's model might want to say that examples like this simply show the need to relax the assumption that successful communication is always worth the same. The payoff for one possible disambiguation could be significantly higher than the other, so that the criterion for successful communication would be satisfied even when the two readings have the same probabilities or accessibility. However, example (9) is specifically constructed to block this kind of explanation. Relevant implications are certainly derivable from Johnny's mother having said that Johnny has a cold, but other relevant implications are derivable from Johnny's mother having said that *Billy* has a cold: for example, that she is forbidding Johnny to go swimming with Billy since people who go swimming with people who have colds are likely to catch cold themselves. The proposed modification of Parikh's model will not account for the data in this case.

It seems to me that there are at least three different lines a defender of Parikh's model might take at this stage. The first is to stick with the original proposal, that ρ and ρ' come from the probabilities of the propositions p and p' but that these probabilities are overridden in special cases, such as (6) and (8). The weaknesses of this account are that it claims that the apparent symmetry of examples like (5) and (6) is only superficial, and, more seriously, that it owes an account of 'special' circumstances and how they affect ρ and ρ'.

The second and third lines of explanation are not mutually exclusive; indeed an explanation of (9) in relevance theory would make use of both of them. The second line is the one suggested above, that ρ and ρ' reflect accessibility but with the further (very reasonable) claim that in cases like (9) the accessibility data from current theories are not complete and that correct measurements of accessibility would include a measure of accessibility-in-the-discourse. This more sophisticated measure of accessibility would make Johnny more accessible than Billy as a referent in (9), regardless of which was subject and which object in the preceding sentence. Something of this sort is probably correct, but I doubt that it is the whole story here.

The third line of explanation seems to me to be at least part of the correct account for examples like (9). In this example the accessibility of background assumptions seems to be crucial. If Johnny's father can access the background assumption (10) then he can make sense of his wife's utterance, taking 'he' to refer to Johnny.

(10) Children who have a cold should not go swimming

As I said above, to make sense of the utterance with 'he' referring to Billy, Johnny's father might access a background assumption like 'People who go swimming with people who have colds often catch cold themselves.' There are other possibilities, but it seems likely that none of them would typically be as accessible as (10).[8] An advocate of Parikh's theory could try to factor in the accessibility of contexts as a contribution to effort, with greater effort reducing the payoff of a particular reading. This would be a considerable step beyond what Parikh considers as effort factors, namely linguistic complexity only, in the case of explicit meaning, and linguistic complexity and the effort involved in forming a mental representation in the case of implicatures. This move would effectively mean adopting the relevance theoretic account of processing effort, which I discuss in section 6 below.

## 5 Implicatures

Consider a case where there is an utterance φ, which could have an implicature but need not.

The total meaning conveyed without the implicature is l; the meaning conveyed with the implicature is p. For example:

(11) φ: "It's 4pm."
l: The time now is 4pm.
p: The time now is 4pm. + It's time to go to the talk.

As before, it is assumed that information is valuable and misinformation has a negative value (of –2 in this case).

It is also assumed that φ is uttered in a context where the speaker knows that the hearer wants to go to a talk at 4pm. In this case p is worth more to the hearer than l, because the extra information would help him in a decision he has to make. In other cases of communication involving implicatures, p may be worth more than l for other reasons (I comment on this in section 6 below). Here the values of p and l are set to 9 and 4 respectively for both speaker and hearer.

On the effort side, it is assumed that there are processing costs for the hearer and for the speaker: they are greater for the speaker because she has to produce the utterance. Parikh also assumes that arriving at p is more costly than arriving at l

[8] Although one candidate for consideration would be an interpretation using the assumption in (10) which would yield the implication 'Billy shouldn't go swimming' hence 'Johnny can't go swimming with Billy', hence a refusal of Johnny's request. (Wilson, p.c.) I suspect that this interpretation as a whole is less accessible than the one reached.

only, since "p has to be contextually inferred from l… [with] additional processing involved." (Parikh, 2001, p. 81)

Then we have the following game:

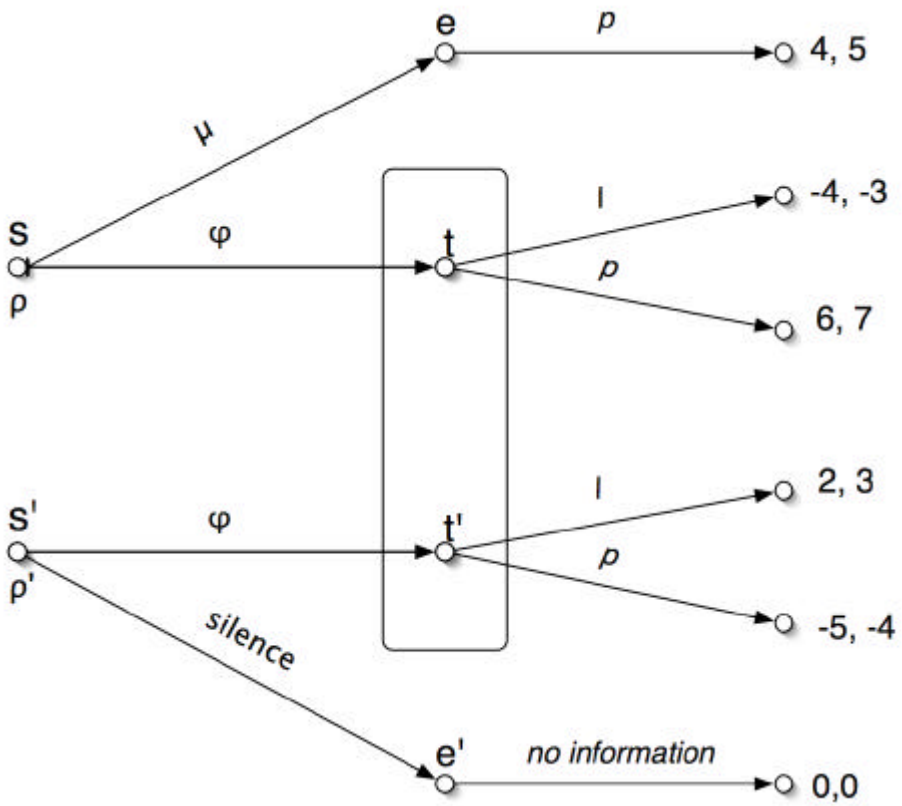


Figure 8 Local game for implicatures (Parikh, 2001, p. 82)

As in the game for explicit meaning, alternative unambiguous utterances are considered, here μ, which unambiguously means p, and silence, which is taken to convey no information, problematically, since silence is often communicative, in fact.

Provisionally accepting all of these assumptions and taking the probabilities of s and s' to be negligibly different, we have as a unique Pareto-Nash equilibrium the strategy in which the speaker utters φ if she wants to convey p and silence if she wants to convey an 'empty interpretation' and in either case is correctly understood by the hearer. How adequate is this kind of account of implicatures? In the next section, I consider some questions which any account of implicatures should answer.

## 5.1 Questions about implicatures

A crucial question for an account of implicatures is: What is the search space for implicatures?

In Parikh's model no constraints are placed on p except that it and l are both meanings of φ within the language that the speaker and the hearer speak. Formally, this is expressed as a meaning function that maps from utterances onto (multiple) meanings. (This was used earlier, implicitly, to give the two possible meanings of example (1).)

Now I think that it is nearly possible to set up the situation this way for explicit meaning, but I am not sure how informative it is to say for implicatures that *a language* would specify all possible meanings in all possible contexts. In my opinion, the reason it might seem right for disambiguation is that there really is a small set of possibilities that are mentally represented—in semantic information associated with lexical items that are present in the utterance. For implicatures, on the other hand, the set of possible meanings would itself depend heavily on context, and Parikh's model does not explain how.

It is unclear whether, and if so why, Parikh wants to abandon the Gricean idea that implicatures should follow logically from the explicit content and some other assumptions. That assumption is desirable because it vastly reduces the search space for implicatures, particularly in the relevance theoretic formulation: implicatures follow from explicit meaning together with contextual assumptions. In Parikh's model, it seems that context places no restriction on the meanings that are possible for an utterance, rather, it is used to guide the selection between possible meanings. Perhaps the reason for this is that in Parikh's model only possible meanings of the utterance are represented and not other propositions which are involved in comprehension. In particular, he does not mention contextual assumptions, so he cannot show how literal meaning logically constrains implicatures. For example, the extra meaning in p over the meaning in l does not follow from l, but from l plus contextual assumptions.

To restate the point: getting disambiguation right can depend on getting the right context. With implicit meaning this is even more clearly important, because there is an open-ended number of candidates to consider for implicatures. That makes it an even more attractive possibility to have a way of getting to the right interpretation that starts from the context in finding candidates to evaluate. (Below I show how this is done in relevance theory.)

These considerations are linked to the next question I want to consider: How is the context for interpretation selected?

I agree with Sperber and Wilson (1986/95, pp. 137-42) that context selection is a very serious problem for pragmatics, if not *the* problem. Along with many other models of communication, Parikh's seems to me to be making use of intuitions of

relevance at this point, at least in the examples given. In the example Parikh chooses p (minus l) as a candidate for an implicature. Intuitively this is a good candidate in this example: it would make sense in the context which is assumed. These intuitive assumptions seem to be taking the place of an account of the way that the hearer decides in what context he should process the utterance to make the intended sense of it. Without wishing to go too deeply into this question here, I want to stress that this is a non-trivial problem. To take an extreme example, Sherlock Holmes may utter a sentence where reference assignment or disambiguation is needed, such as "He went to the bank," and his hearer, Dr. Watson, may be simply unable to work out which person and what kind of bank are involved because although he and Holmes are in the same physical environment, perhaps even attending to the same aspects of it, and have heard the same prior discourse, still he and Holmes are in different *contexts*. This is because Watson has not made the same inferences as Holmes and thus starts with different information and—importantly—because he is simply unable to make those inferences and therefore cannot get himself into the same context as Holmes. [9]

A good account of communication need not be a general theory of contexts, of course, but it must have something to say about the way in which a hearer, presented with an utterance, searches for a context which the speaker took to be accessible to him and in which the utterance can be interpreted so as to make the kind of sense the speaker intended. I am not aware of any consideration of these issues by game theorists interested in communication.

Note that the contextual assumptions involved do not need to be things that the hearer believed or had even thought about before the utterance, as Sperber and Wilson's example (12) shows:

(12) Mary and Peter are looking at a scene which has many features, including what she knows to be a distant church.
Mary: I've been inside that church. (Sperber & Wilson, 1986/95, p. 43)

> Mary "does not stop to ask herself whether he has noticed the building… All she needs is reasonable confidence that he will be able to identify the building as a church when required to… it might only be on the strength

[9] I am thinking of cases such as the following from *The Blue Carbuncle*:

> "I can see nothing," said I, handing [a hat] back to my friend.
> "On the contrary, Watson, you can see everything. You fail, however, to reason from what you see. You are too timid in drawing your inferences."(Conan Doyle, 1897/1957, p. 125)

> of her utterance that it becomes manifest to him that the building is a church." (ibid., pp. 43 f.)

Sperber and Wilson develop an account, including the notion of mutual manifestness, which explains examples like these. A proponent of Parikh's account would need to show that it could also deal with context selection.

A further problem for modelling any open-ended inference process is proposing a stopping rule that works. My final question in this section, then is: How is the search stopped?

Parikh gives an example of the way his model stops further implicatures from being generated. I would like to suggest that to adopt his answer is effectively to adopt a principle of relevance.

The example considered (Parikh, 2001, pp. 84f) models a situation where there are three possible interpretations to be considered if the speaker utters φ: l, p and q. l and p are as before; q includes p plus some extra information. The result is a more complicated game with a solution consistent with the previous one—the speaker, wishing to convey p, utters φ and the hearer interprets it as p—as long as the value of the new information is less than the effort expended in processing it. This follows because choosing p will be the equilibrium strategy for the hearer (after φ is uttered) when the payoff for choosing q rather than p is lower because the cost of the new information outweighs its value. If q is p plus something informative but irrelevant, as in Parikh's example, "let's take the route by the lake", then as Parikh puts it: "Suppose there is some other proposition, q, ... that is more informative than p. It will certainly have positive value but it is easy to see that it cannot have more value than p [in the context discussed]. ... Moreover it is reasonable to assume that the greater the information, the more costly it is to process." (2001, p. 84).

Effectively, then, the stopping criterion is: generate implicatures until the effort involved in doing so is less than the value of the information. This is close to the communicative principle of relevance, discussed below. If Parikh's model needs this principle and relevance theory can provide an account of communication with a similar principle and little other machinery then Parikh's account seems to suffer from relative lack of economy. Parikh's principle of relevance may differ significantly, however, from the relevance-theoretic principle of relevance in that it seems to look for maximal rather than optimal relevance. If so, the two different principles would make some different predictions. I discuss this in the next section which introduces relevance theory and compares relevance theory and Parikh's model at several points.

## 6 Relevance theory and communication

Relevance theory is a theory of cognition. (Sperber & Wilson, 1986/95; Sperber & Wilson, 2002) It claims that human cognition tends to be geared to the maximization of relevance. (This is the cognitive principle of relevance). Relevance is defined in terms of cognitive effects and processing effort:

(13) *Relevance*
   a. The greater the cognitive effects, the greater the relevance;
   b. The smaller the effort needed to achieve those effects, the greater the relevance.

Cognitive effects occur when new information interacts with existing contextual assumptions in one of three ways:

(14) *Cognitive effects*
   a. Strengthening an existing assumption;
   b. Contradicting and eliminating an existing assumption; or
   c. Combining with an existing assumption to yield contextual implications.

(15) *Processing effort* is affected by:
   a. the form in which the information is presented
   b. the accessibility of the context.

In the special case of ostensive inferential communication, the speaker, by making an utterance, is making an offer of information. This raises the expectation that the utterance will be optimally relevant:

(16) *Optimal relevance*
   An utterance is optimally relevant to an addressee iff:
   a. it is relevant enough to be worth the addressee's processing effort;
   b. it is the most relevant one compatible with the speaker's abilities and preferences.

This expectation is spelled out in the Communicative Principle of Relevance:

(17) *Communicative Principle of Relevance*
   Every utterance communicates a presumption of its own optimal relevance.

This implies the relevance-theoretic comprehension procedure:

(18) *The relevance-theoretic comprehension procedure*
   a. consider interpretations in their order of accessibility (i.e. follow a path of least effort);
   b. stop when the expected level of relevance is achieved.

Compare this with Parikh's account of communication. Parikh also factors in effects and effort, but he is less precise about what may contribute to them. Effects enter the calculation as the 'value of information' which contributes positively to payoffs. As mentioned, an initial, provisional assumption is that all information is equally valuable; later Parikh shows how information can be assigned a value by considering its worth in a game modelling a decision that (the speaker knows) the hearer may make. As with Lewis' work this seems to blur the distinction between the illocutionary and perlocutionary aspects of meaning. Parikh recognizes that there is an issue here, and separates implicatures into two types: type I, where an utterance has a direct effect on the hearer's behaviour and type II, where an utterance affects only the hearer's thoughts, initially at least. In the second case Parikh says, "this type of implicatures can be modeled in more or less the same way except that we need to consider preferences for information directly, rather than via direct action." (2001, p. 86) In Parikh's type II examples, the estimation of the value of the information in these cases is apparently arrived at intuitively. Parikh's model does not specify as relevance theory does in (14) the ways in which information can be valuable to the hearer. Someone wanting to make the model more precise would have to consider whether to adopt the relevance theoretic postulate, that information can be valuable by reinforcing, contradicting or combining with existing information, or make some other specification.

Another problem is the division of implicatures into two types. I think that this is undesirable for two reasons. First, it seems that explicatures[10] as well as implicatures sometimes lead more directly to action than at other times. Should there also be two categories of explicature? Secondly, the first type of implicature seems redundant. Presumably all utterances, including ones that lead fairly immediately to decisions and actions, have their effects on the hearer by affecting his thoughts. (Or at least, in cases where this is not true we would not want to call what has happened communication.) So all implicatures will belong to Parikh's second type.

I have already commented that Parikh allows as effort factors only linguistic complexity (with the metric unspecified) and, later, the cost of representing or

[10] In relevance theory the explicitly communicated meaning of an utterance is made up of one or more explicatures, so called in contrast to *im*plicatures. Carston defines an explicature as "an ostensively communicated assumption which is inferentially developed from one of the incomplete conceptual representations (logical forms) encoded by the utterance." (2002, p. 377)

processing an implicature mentally. I have argued that in order to account for examples such as (4) to (9), the model needs to incorporate (15) from relevance theory, so that effort is partly due to the accessibility factors connected with linguistic items in the utterance and partly to the accessibility of contexts.

There is another issue connected with effort which presents central problems for a game theoretic account of communication. In game theory the payoffs do not include the cost of constructing the representation of the game, understanding it and finding a solution or solutions. Parikh's model sets up all of the possible meanings of the utterance in parallel and therefore shares a problem with truth-based approaches. As Wilson and Matsui (1998) point out, in order to find which interpretation is most likely to be true, all possible interpretations must be considered. This makes these approaches psychologically implausible. Parikh's model has this problem doubled or quadrupled. First, both the speaker and the hearer must consider all possible interpretations; secondly, for every interpretation an unambiguous alternative utterance must be found.[11] Considering all possible interpretations is not a trivial matter even for explicit meaning—Parikh's example (1) has at least eight different possible readings since there are at least three degrees of freedom, given that the meaning of 'New York' and the precision or otherwise of 'ten minutes' as well as the scope of the quantifiers are underdetermined by the linguistic form. For implicatures there does not seem to be any principled reason why there should be a determinate number at all; at the least, there must be a huge finite number of interpretations that any given utterance can have. In fact this kind of indeterminacy also gets into the explicit meaning, as for example in (1) where resolution of the meaning of 'ten minutes' is more a matter of finding a degree of precision on a continuum than of choosing among a limited number of possibilities. Contrast Parikh's model with the relevance theoretic comprehension procedure which:

> integrates effort and effect in the following way. It claims that the hearer is entitled to expect at least enough cognitive effects to make the utterance worth his attention, that the processing effort is the effort needed to achieve these effects, and that the hearer is entitled to accept the first interpretation that satisfies his expectation of relevance. (Wilson and Matsui, 1998, p. 18)

---

[11] I do not want to deny that consideration of alternative utterances often plays a role in hearers' recovery of meaning (and speakers' choices of wording). It seems that it would be good for a theory only to take alternative utterances into account where they are easily accessible and to use at most the fact that they are *not* easily accessible in other cases. This is how relevance theory deals with these cases (eg Sperber & Wilson, 1986/95, pp. 200-1.)

So the hearer works through interpretations in order of (decreasing) accessibility until one of them has cognitive effects worth the processing effort—in which case this will be the interpretation understood (or until the pragmatic faculty exceeds the amount of effort it can spend on this occasion and gives up—in which case no interpretation will be arrived at). In practice this means that generally not very many interpretations will need to be considered; often, as in (5) and (6), the most accessible interpretation will be the right one. Thus while any interpretation might be considered, combinatorial explosion is avoided. The relevance theoretic comprehension procedure is simple and seems computationally tractable: it is a fast and frugal heuristic, in Gigerenzer's terms (Sperber & Wilson, 2002, section 5).

Parikh can claim that his model does not need to answer the charge of psychological implausibility since it may be that the model does not describe cognitive structures or processes. As he says, "It seems better to view the game as a model of a class of constraints that capture the underlying logic of communication… . The game …describes a valid inference without saying anything about how agents arrive at the correct interpretation of an utterance." (Parikh, 2001, p. 25) I do not think this is quite right. Certainly, comprehension might be carried out by a heuristic which arrives at the same interpretations as the model (at least often enough). But if the model correctly describes the logic of the situation then it implies that the mental processes involved in communication, if communication occurs at all, must be sufficiently sophisticated to grasp the situation correctly in some way. So the more complicated the model of the situation, the more mysterious the success of the heuristic and the more difficult it would seem to give an account of the workings of that heuristic.[12] In other words, the model does say something about 'how agents arrive at the correct interpretation' at least by specifying the nature and complexity of the problem that they have to solve. Arguably, an account of communication which shows that communication is complicated without making any suggestions about how people manage to understand each other is lacking in a crucial respect.[13]

---

[12] This argument is a distant relative of an argument used recently in minimalist syntax. (For example by Tanya Reinhart at ESPP 2003, commenting on Chomsky, 2001.) It is justified to include effort considerations in our account of syntactic competence, according to the argument, because the parser has to deal with the output of the competence and, on the face of it, the more complicated are the representations the competence generates, the more difficult is the job of the parser.

[13] One response to this criticism might be that only pragmatics needs to say anything about *how* communication is achieved. Parikh explicitly denies that his account is a pragmatic theory: "I see this whole book as a part of semantics, not pragmatics, because rational agency is part of a properly situated semantics." (2001, p. 141) I do not find this line of argument very impressive. Both pragmatic theories and Parikh's model are in the business of explaining communication. If a particular pragmatic theory can say *what* is communicated and *how* the hearer recovers the

Recall that in section 5.1 above I argued that Parikh's method of stopping in implicature derivations effectively relied on something very like a principle of relevance, that is, something like: keep generating implicatures as long as the value of the information in an implicature is greater than the effort costs associated with it. This seems to be a principle of maximal relevance, since it makes search continue while there is any more value to be obtained that is worth the effort used in obtaining it. In contrast, the principle of communicative relevance claims that search will only continue until an optimally relevant interpretation is reached. The difference is that relevance theory claims that hearers look for an interpretation such that the utterance was "the most relevant one *compatible with the speaker's abilities and preferences*." (my emphasis) A theory which claims that hearers look for maximal relevance, ignoring these provisos, predicts certain implicatures that do not arise in relevance theory. Carston discusses a number of cases like this in a paper on so-called 'scalar implicatures' (Carston, 1998) including example (19) (her example (71), taken from Green (1995, 96-97):

(19) B: Are some of your friends Buddhist?
A: Yes, some of them are.

Theories which claim hearers look for maximal relevance, including Parikh's model apparently, predict here that A will be taken to implicate that not all of her friends are Buddhist, since "in the context that Green sketches, it is evident that there is a more relevant response that A could have given, concerning whether all or most of her friends are Buddhist; this would have more contextual effects for the hearer (B) and would cost him negligible further processing effort Since A has chosen not to utter this, doesn't it follow that she must be communicating that *only some* (that is, not all or most) of her friends are Buddhist?" (Carston, 1998, p. 33). On the other hand, relevance theory correctly predicts that this implicature will not arise if it is manifest that the speaker was not able or not willing to make a stronger statement. "Green's context makes it plain that while the speaker has the *ability* to make the stronger statement, she *prefers* not to (she is afraid of being considered a Buddhist-groupie) and the hearer is aware of this. Hence the relevance principle correctly predicts that the speaker is not implicating that not all of her friends are Buddhist and that the hearer recovers no such assumption as part of what is communicated."(ibid.)

However, Parikh's model may not be in as much trouble with this kind of example as other frameworks in which relevance is effectively maximised. Many neo-Gricean and post-Gricean accounts of communication assume as a

---

intended meaning and Parikh's model only has an answer to the first question, then *ceteris paribus* the pragmatic theory is to be preferred.

foundational principle, as did Grice, that communication is cooperative. As a consequence they are simply unable to give an account of utterances where the speaker will not cooperate. Parikh's model does not assume a Cooperative Principle, so, at least in principle, it could make correct predictions in these cases, given assumptions about the way the speaker's preferences affect payoffs. As far as I can see, a defender of Parikh's model would have to write in a proviso like the second half of the second clause of optimal relevance, so that the value of extra implicatures would be zero, no matter how useful the information might be to the speaker, if the speaker manifestly was not willing or able to communicate them.

I have another worry, however, about the optimization of interpretations in Parikh's account. According to relevance theory, when hearers try to find the interpretation of an utterance intended by the speaker there does not have to be any cooperation between the speaker and hearer, except that the speaker wants to be understood and the hearer to understand (Sperber & Wilson, 1986/95, p. 268). The speaker knows that the hearer is built so as to take the first interpretation that is optimally relevant as the correct one, so she has to make sure that her utterance will lead the hearer to entertain this interpretation before any other which would be relevant. Parikh's model has a similar asymmetry between the speaker and the hearer: both must consider the local game which determines the interpretation of an utterance but only the speaker needs to consider the global game, choosing the utterance which has the highest payoff given an intended interpretation. There is a difference, however. According to relevance theory the interpretation must be optimally relevant for the hearer but not, in general, for the speaker. The constraint from the speaker's point of view is to produce an utterance which is optimally relevant to the hearer, compatible with the speaker's abilities and preferences. In contrast, in Parikh's model the solution must be optimal for both speaker and hearer for successful communication. There seems to be something problematic about this, since the reasons why the solution will be optimal will generally be different for the speaker and the hearer. Which interpretation will be optimal for the hearer depends on the worth to him of the information he can derive from it. For the speaker the optimal solution is simply the one in which the hearer arrives at the interpretation the speaker intended (or something close enough to it). So the payoffs for the speaker and the hearer will not generally be the same. While Parikh's model allows for this, so it is not a problem in itself, there seems to be a worry here, since miscommunication will occur if the payoffs come apart too far and a defender of the model would need to show that this does not generally happen in ordinary cases where, as we have seen, generally the interests of the speaker and hearer are different.

This is one aspect of a general worry about the model, since Parikh allows that a number of factors—the effort and effect factors in the payoffs, the probabilities and even the set of meanings for an utterance—can be different for the speaker and

hearer. Any of these might come apart, perhaps leading to miscommunication. To consider just the effort factor, this will generally be very different for the speaker and the hearer even though the psycholinguistic costs of an utterance are often assumed to be the same. Other costs may well be different since the tasks involved are different. The speaker knows (roughly) what she wants to mean, and has to work out what utterance will suit; the hearer knows what has been uttered and has to work out what was meant, including implicatures—which may require considerable inference.

## 7 Rationality and game theory

### 7.1 The other type of model

In section 2.1 I noted that van Rooy (forthcoming) applies game theory to communication in a different way from Parikh, taking rationality and economy considerations to apply to language rather than utterances in context. Nonetheless there are similarities between the models. Van Rooy looks at a situation where there are two different possible meanings, one more salient than the other, or 'unmarked', and two utterance-types, one less linguistically complex than the other (or 'unmarked'). He, like Parikh, finds two Nash equilibria. In his model these represent possible linguistic conventions. N1, the intuitively correct solution, is the convention corresponding to Horn's 'division of pragmatic labour': unmarked utterances carry unmarked meaning; marked utterances carry marked meaning. The other Nash equilibrium, N2, is the 'anti-Horn' case: unmarked utterances carry marked meaning; marked utterances carry unmarked meaning. Van Rooy rejects Pareto-dominance as a secondary criterion for equilibria, so both N1 and N2 are solutions, which implies that both the Horn and the anti-Horn case are conventions in Lewis' sense (Lewis, 1969)—they are non-unique stable solutions, so that adherence to one or the other is arbitrary to some degree. This suggests that some speech communities should follow the Horn convention and others the anti-Horn convention. This seems to me to be an unfortunate result. I doubt whether there has ever been a speech community in which marked utterances typically have unmarked meanings and unmarked utterances typically have marked meanings. In no community would an utterance of "Miss X produced a series of sounds that corresponded closely with the score of 'Home sweet home'" (Grice, 1967, p. 54) mean simply that Miss X sang 'Home sweet home' where a sentence "Miss X sang 'Home sweet home'" could be uttered.

Van Rooy (forthcoming, section 5) suggests that it may be possible to eliminate N2 by showing that only N1 is evolutionarily stable.[14] Thus the existence of the two equilibria could be reconciled with the non-existence of anti-Horn communities. This strikes me as a very convoluted way of arriving at a result that, to the degree that it is true, falls out naturally from the communicative principle of relevance: if a speaker puts a hearer to more trouble than he might have (for example by using a 'marked' utterance), then the hearer is entitled to more cognitive effects (in other words, a 'marked' interpretation).

I think that there are also more general problems with the approach which applies game theory to language rather than to particular utterances. It seems to confuse language with communication, as for example when van Rooy writes that in this approach, "Speakers obey Horn's rule because they use a conventional language that, perhaps due to evolutionary forces, is designed to minimize the average effort of speakers and hearers", apparently neglecting considerable evidence that languages are not particularly economical considered as tools for communication, for example that syntactic competence may generate sentences that are unusable for communication or perhaps for any purpose. Taking language and communication systems to be separate allows a more insightful approach to understanding communication—communication systems use whatever means are available, sometimes language, but far from always.

## 7.2 Empirical and theoretical doubts about game theory's formalisation of rationality

In section 2.1, I mentioned that empirical and theoretical questions have been raised about the way that game theory formalizes rationality. (See Colman, in press, for a summary.) Since a game theoretic account of human communication would inherit these problems, it seems relevant to give a brief sketch of some examples. I will mention one empirical and one theoretical worry.

Figure 9 shows a four-legged version of Centipede (Rosenthal, 1981 and Binmore, 1987, cited in Colman, in press). Player A and player B make choices between cooperating, which continues the game to the right, and defecting down, ending the game. Every time a player cooperates, his payoff if the game ends at the next decision node is less by 1 than he would have earned from defecting immediately and his opponent's payoff increases by 10.

How should a rational agent play this game? At the final decision node B should defect down to receive 19 rather than the 18 he would receive by continuing across. A knows this so should defect down at the previous node to receive 9 rather than 8.

---

[14] Similarly, Asher et al (1999) attempt to derive a Gricean maxim of truthfulness using evolutionary game theory.

Similarly, B should defect down at the previous node to receive 10 rather than 9, so we arrive at the result that A should defect down at the first node to get 0 rather than –1. This kind of argument shows, amazingly, that in Centipede the first player should always defect down at the first node—with both players receiving zero—no matter how long the Centipede is and how high are the final payoffs.

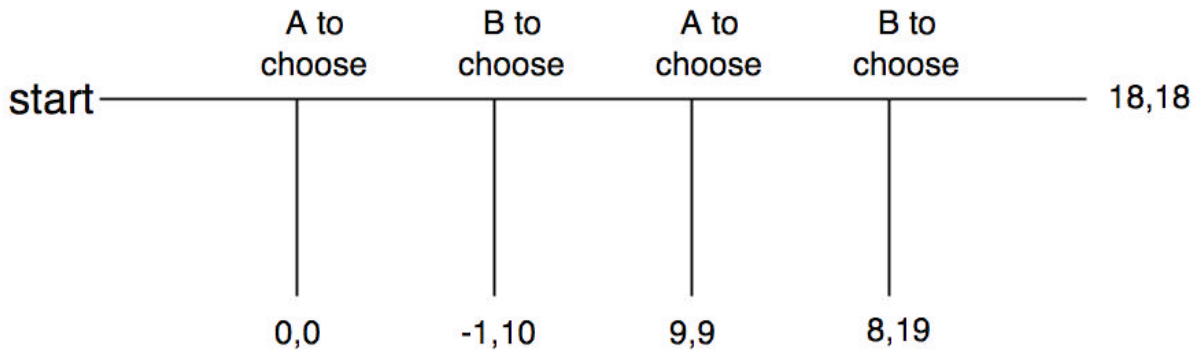


Figure 9 A four-legged version of Centipede (after Colman, in press, section 7.3)

The experimental evidence is that players are considerably more cooperative, "the great majority continuing across at the first decision node, and a substantial minority even at the last." (Colman, in press, section 7.4) Now game theory is primarily normative rather than positive so it is not necessarily a problem for game theory if people behave differently from game-theoretic rational agents. However it does seem to pose a problem to game theory that people do systematically better than such agents in certain games. As Colman says, "cooperative players, who earn significant monetary payoffs in experimental Centipede games, could legitimately ask rational utility maximizers, who come away with nothing: "If you're so rational, how come you ain't rich?" This seems a good question, given that rational decision making is, by definition [in game theory], utility maximization." (ibid.) This suggests that there may be something wrong with the formalisation of rationality within game theory.

A further worry is also illustrated by Centipede games: the rationality assumptions of game theory may be incoherent. Certainly there appears to be no way to avoid inconsistency in applying them to Centipede games. Consider A's decision at the penultimate node. To rule out the path leading to payoffs of 18,18, he needs to make use of the assumption that B will act rationally by maximizing his utility at the final node. But for the game to have reached the penultimate node, B must have chosen not to defect at the previous node, apparently irrationally. This seems to mean that B is not predictably rational, so A cannot know what B will do at the final node. In this situation neither player can act rationally, it seems. (Colman, in press, section 7.5).

I have illustrated only two of the problems summarized thus by Colman: "Instrumental rationality, conventionally interpreted, fails to explain intuitively obvious features of human interaction, yields predictions starkly at variance with experimental findings, and breaks down completely in certain cases (ibid, abstract.)" Until there is greater understanding of these issues, perhaps within *psychological* game theory (ibid.) it will be unclear whether a game-theoretic account of communication successfully links communication with reasonable assumptions about rationality.

One assumption made by game theory, that the players have common knowledge of the game and of the rationality of the other player, may prove particularly problematic for game-theoretic accounts of communication. Sperber and Wilson discuss reasons why common knowledge of *context* is not a reasonable assumption for a pragmatic theory (1986/95, pp. 15-21). There need not be a direct clash between these positions (pace Sally, 2003, p. 1234), but it may be that game theorists also wish to build in common knowledge of (some) features of the context. Further, some of Sperber and Wilson's arguments may be effective against common knowledge of the game and of rationality assumptions, in which case any treatment of communication employing standard game theory will be rendered doubtful. Contra Sally, there is no presumption in favour of common knowledge and game theory here, certainly not before there is a fairly successful game theoretic account of communication—whose success might then be taken as corroboration of its assumptions.

## 8 Conclusions

In this paper I have tried to explain and explore Prashant Parikh's game-theoretic model of communication in order to illuminate the possibilities that game theory offers for understanding communication. I have raised empirical and theoretical doubts about this model and, briefly, about van Rooy's alternative account, and reviewed some questions about the foundations of game theory.

The problems for Parikh's account that I have examined fall into three categories. I have expressed doubts about its coverage of data; claimed that it relies on intuitions about relevance; and raised questions about its implications for a psychologically plausible account of communication.

I believe that there is another kind of problem faced by the theory. It seems to me to be the wrong kind of theory of communication, in that it models communication as the transfer of thoughts between the mind of the speaker and the hearer, reducing the role of inference in communication to a way of implementing coding and decoding. This is a common way of reconciling Grice's insight that communication is inferential with a widespread folk theory of communication: that it is a matter of

getting across a 'signal' or a 'message'. (Advocated by Searle, for example, according to Sperber & Wilson, 1986/95, p. 25). Sperber and Wilson argue against this move, pointing out that communication is possible in the absence of a code (e.g. Mary shows Peter a bottle of aspirin in response to his question 'How are you feeling today?'(ibid.)) and that only a truly inferential account can deal in a unified way with examples like this as well as cases where utterances include conventional linguistic or gestural meaning. In the introduction I claimed that 'transfer' models of communication are connected to the notions of expressivity of language or effability of thoughts. If every thought had a pronounceable linguistic version, then the role of inference in communication would be limited to resolving ambiguities that arise only because the unambiguous statements of thought are uneconomical to utter or understand. There are good reasons to distrust the expressivity thesis, among them that thoughts are probably indexical to the thinker, but not least that we seem to be able to infer from utterances (and think) thoughts that go well beyond what the words and phrases we utter seem to mean. There are prosaic examples: the degree of tiredness conveyed by a given utterance of 'I'm tired' depends on the context, with many different kinds of tiredness thinkable but not lexicalized (Carston, 2003, p. 40); as well as examples of poetic effects, it being well known that no good poem (or joke) can be exactly paraphrased.

I would be more hopeful, then, about the success of future game-theoretic accounts of communication if they adopt the null hypotheses that thought is quite distinct from language and that communication essentially involves inference.

## References

Asher, N., Sher, I., & Williams, M. Game theoretical foundations for Gricean constraints. *Proceedings of the Thirteenth Amsterdam Colloquium, ILLC, Amsterdam*.

Breheny, R. (1999). Context Dependence and Procedural Meaning: The Semantics of Definites. PhD diss. University College London.

Carston, R. (1998) Information, Relevance and Scalar Implicature. In Carston and Uchida, eds. Relevance Theory: Applications and Implications. Amsterdam: John Benjamins.

Carston, R. (2002) Thoughts and Utterances: The Pragmatics of Explicit Communication. London: Blackwell.

Conan Doyle, A. (1892/1957) The Blue Carbuncle. In *The Adventures of Sherlock Holmes*. References are to the 1957 edition. London: John Murray.

Chomsky, N. (2001) Beyond Explanatory Adequacy. Ms., MIT.

Grice, H. P. (1967). Logic and Conversation. In Cole, P.& Morgan, J. (1975) *Syntax and Semantics of Speech Acts.* New York: Academic Press. Also in Grice, H. P. (1989) *Studies in the Way of Words.* Cambridge, MA: Harvard University Press. References here are by page numbers in Cole & Morgan.

Katz, J.J. (1981) *Language and Other Abstract Objects.* Oxford: Basil Blackwell.

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

Parikh, P. (1991). Communication and strategic inference. *Linguistics and Philosophy, 14*, 473-513.

Parikh, P. (2000). Communication, meaning and interpretation. *Linguistics and Philosophy, 23*, 185-212.

Parikh, P. (2001). *The Use of Language.* Stanford, California: CSLI Publications.

Sally, D. (2003). Risky speech: behavioral game theory and pragmatics. *Journal of Pragmatics, 35*, 1223-1245.

Sperber, D., & Wilson, D. (1986/95). *Relevance: Communication and Cognition (2nd edition)*. Oxford: Blackwell.

Sperber, D., & Wilson, D. (1997/98). The Mapping Between the Mental and the Public Lexicon. *UCL Working Papers in Linguistics, 9.* 107-25 Reprinted in Carruthers, P. & Boucher, J. (eds.) (1998) *Language and Thought: Interdisciplinary Themes,* 184-200. Cambridge: Cambridge University Press.

Sperber, D. & Wilson, D. (2002/ in press) *Relevance Theory.* To appear in Ward, G. & Horn, L. (eds.) Handbook of Pragmatics. Oxford: Blackwell.

van Rooy, R. (in press). Signaling games select Horn strategies. *Linguistics and Philosophy,* forthcoming.

van Rooy, R. (to appear). Being polite is a handicap: Towards a game theoretical analysis of polite linguistic behaviour. In the proceedings of TARK 9.

Wilson, D & Matsui, T. (1998) Recent Approaches to Bridging: Truth, coherence, relevance. *UCL Working Papers in Linguistics 10. 173-200*.