Signalling games select Horn strategies

Robert van Rooy*

Abstract

In this paper I will discuss why (un) marked expressions typically get an (un)marked interpretation: Horn's division of pragmatic labor. It is argued that it is a *conventional* fact that we use language this way. This convention will be explained in terms of the equilibria of *signalling games* introduced by Lewis (1969), but now in an *evolutionary* setting. I will also relate this signalling game analysis with Parikh's (1991, 2000, 2001) game-theoretical analysis of successful communication, which in turn is compared with Blutner's (2000) bi-directional optimality theory.

1 Introduction

Yesterday, Paul came into my office and told me 'Miss X produced a series of sounds that corresponds closely with the score of "Home Sweet Home"; Paul intended to communicate something to me and he succeeded: I understood that Paul wanted to tell me that Miss X's performance suffered from some hideous defect. How can this be explained?

The above example is just one instance of a general rule that says that (un)marked expressions typically get an (un)marked interpretation. Many other examples are discussed in Horn (1984) and the rule has come to be known as *Horn's division of pragmatic labor*. Sometimes I will also denote this rule by *Horn's rule* or the *Horn strategy*. Grice (1975) appeals to his maxim of *manner* to explain the example I started out with. He also suggests

*The research for this paper has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences (KNAW). I would like to thank Boudewijn de Bruin, Paul Dekker, Jason Noble, Brian Skyrms, Frank Veltman and Henk Zeevat for discussion. Special thanks, however, I owe to Johan van Benthem, Reinhard Blutner, Prashant Parikh, Katrin Schulz, Sarah Taub and two anonymous reviewers for L&P: most of the changes from initial manuscript to final paper were triggered by their comments and inspiring questions. This doesn't necessarily mean, of course, that they agree with everything I claim. I thank William Rose for correcting my English.

that this maxim, just like the other ones, should be derivable from general principles of rationality. This seems natural: when we obey the rule, we use linguistic expressions in a *more economical* way than when we don't. But what should such a reduction to principles of economy and rationality look like?

According to a tradition going back to Zipf (1949), economy considerations apply in the first place to *languages*. 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. Horn's (1984) own explanation in terms of the interaction of his Q and R principles belongs to this tradition,[1] and so does the recent Optimality Theoretic one of Blutner (2000).

According to another possible way to go, rationality considerations apply every time a *speaker* and a *hearer* are involved in communication. The 'rule' that Horn observed is not a convention among language users, but is observed only because rationality dictates that speaker and hearer always coordinate their utterance and interpretation acts in accordance with the rule. Parikh's (1991, 2000, 2001) game-theoretical explanation of successful communication is perhaps the most explicit analysis following this road.[2]

The main goal of this paper is to convince you that (in general) the first line of attack is more natural than the second. To do so, I will give a game-theoretical explanation of how Horn's rule could have become conventionalized through the forces of evolution. But the paper has some secondary goals too: (i) to show the similarity of Blutner's and Parikh's analyses of successful communication; (ii) to point out the resemblance of Parikh's unusual games of partial information with so-called *signalling games* introduced by Lewis (1969); and (iii) to point out that recent work on signalling games within economics is of interest to the semantic/pragmatic analysis of natural language: perhaps linguistic models that try to account for the interpretation of expressions that are *partly* underspecified by semantic constraints can benefit from economic models that explain the interpretation of signals that have no *a priori* given meaning at all.

This paper is organized as follows: In section 2, I discuss Blutner's bidirectional OT and its game-theoretical reformulation due to Dekker & van Rooy. Parikh's more general game-

[1] See also Atlas & Levinson (1981) and Levinson (2000).

[2] To be fair, Parikh (2001) also mentions the idea that the relevant economic calculations of language users are perhaps *unconscious*, while Blutner, Horn and Zipf sometimes suggest that their analyses are relevant for *particular* conversational situations.

theoretical analysis of successful communication is discussed in section 3 and it is shown how Parikh's analysis accounts for Horn's division of pragmatic labor. Section 4 deals with Lewisian signalling games and with a signalling game reinterpretation of Parikh's analysis. In section 5, I discuss various ways in which Horn's division of pragmatic labor might be given an evolutionary explanation. The paper ends with some conclusions and potential future applications of the framework.

2 Bidirectional OT and Strategic Games

2.1 Bidirectional Optimality Theory

Inspired by Horn's (1984) 'formalization' of Zipf's principles of minimization of speaker's and hearer's effort, Blutner (2000) proposes to account for the phenomenon that (un)marked expressions typically get an (un)marked interpretation in terms of his Bidirectional Optimality Theory. The idea behind Optimality Theory (OT) in semantics/pragmatics (cf. Hendriks & de Hoop, 2001) is that conventional meaning *underspecifies* the actual interpretation of an expression, and that a combination of violable optimality theoretic constraints determines the optimal (= actual) one of those candidate interpretations. The crucial distinction between Blutner's Bi-directional and standard uni-directional OT, is that in the former, but not the latter, for the *hearer* to determine what the optimal interpretation of a given form is, he must also consider the *alternative expressions* the *speaker* could have used to express this meaning/interpretation. One way to implement this idea is to say that we not only require that *the hearer* finds the optimal meaning for a given form, but also that *the speaker* expresses the meaning he wants to communicate by using the optimal form. Thus, what is optimal is not just meanings with respect to forms, but rather form-meaning pairs. Jaeger (2000) connects Blutner's ideas with standard Optimality Theory by showing how the ordering relation between form-meaning pairs can be derived from a system of ranked OT constraints: some of them are relevant only for ordering forms, others only for ordering meanings. Now we can say that form-meaning pair $\langle f, m\rangle$ is **strongly optimal** iff it satisfies both the speaker's principle (S) (i.e. is optimal for the speaker) and the hearer's principle (H) (i.e. is optimal for the hearer):[3]

[3]According to optimality theory there exists also a generation function, G, that assigns to each form f a *set* of interpretations that it could possible mean. For ease of exposition I will ignore this function,

$$(S) \qquad \neg\exists f' : \langle f, m\rangle < \langle f', m\rangle$$
$$(H) \qquad \neg\exists m' : \langle f, m\rangle < \langle f, m'\rangle$$

Bidirectional OT wants to account for the fact that we typically interpret the lighter form as having a more salient, or stereotypical, meaning. Grice's example of 'singing' versus 'producing a series of sounds' with which I started this paper is one concrete example of this. Another one, discussed by McCawley (1978), is that although 'kill' and 'cause to die' could in principle mean the same thing, we typically will interpret the former lexicalized expression as denoting stereotypical killing (by knife or pistol), while the use of the morphologically complex expression suggests that the murderer performed his action in a less conventional way. It is easy to see that Blutner's notion of *strong optimality* can account for one half of this principle. If we assume that $\langle f, m\rangle > \langle f', m\rangle$ iff f is a lighter expression than f' and that $\langle f, m\rangle > \langle f, m'\rangle$ iff m is more salient or stereotypical than m', it immediately follows that 'kill' gets interpreted as stereotypical killing. We are not able yet, however, to explain how the more complex form can have a meaning at all, in particular, why it will be interpreted in a non-stereotypical way. To account for this, Blutner (2000) introduces a weaker notion of optimality. A form-meaning pair $\langle f, m\rangle$ is **weakly-optimal** iff it satisfies both of the following more complex S and H principles (where $\langle f, m\rangle \in H$ iff $\langle f, m\rangle$ satisfies the new (H)):

$$(S) \qquad \neg\exists f' : \langle f', m\rangle \in H \;\&\; \langle f, m\rangle < \langle f', m\rangle$$
$$(H) \qquad \neg\exists m' : \langle f, m'\rangle \in S \;\&\; \langle f, m\rangle < \langle f, m'\rangle$$

Jaeger (2000) notes that although the S and H principles interact with each other, this does not give rise to an infinite regress as long as we assume that the OT constraints generate a well-founded ordering relation on form-meaning pairs.[4] All form-meaning pairs that are strongly optimal are also weakly optimal. However, a pair that is not strongly optimal like ⟨'Cause to die', unstereotypical killing⟩ can still be weakly optimal: although a stereotypical killing would be the optimal meaning for 'Cause to die', this interpretation is

but all form-meaning pair combinations that play a role in the definitions will obey this constraint: for all $\langle f, m\rangle$ mentioned, $m \in G(f)$.

[4]Benz (ms) has argued recently that this assumption is natural only when the OT constraints are context-independent, i.e., when it is common knowledge that there is a single ordering that works for both players.

blocked by the S principle, because this meaning can be expressed by the lighter expression 'kill'. Similarly, an unstereotypical killing cannot be expressed by 'kill' because this is blocked by the H principle: there is a less marked meaning that could be denoted by 'kill'. The pair ⟨'Cause to die', unstereotypical killing⟩ is not blocked at all, however, and thus weakly optimal.

2.2 A game-theoretical reformulation

Blutner's bidirectional OT has been given a game-theoretical reformulation in Dekker & van Rooy (2000). According to this reformulation, information exchange is represented as a strategic (interpretation) game of complete information between speaker and hearer. A two-player **strategic** game, or a game in strategic form, is a model $\langle\{1,2\},(A_i),(U_i)\rangle$ of interactive decision making in which each player i (element of $\{1,2\}$) chooses her plan of action (element of A_i) once and for all, and is uninformed, at the time of her choice, of the other players' choices. The actions chosen by the players depend on their preferences, modeled in terms of a cardinal utility function (U_i) over the action profiles, the simultaneous choices of the players. A profile $\langle a_1, a_2\rangle \in A_1 \times A_2$ of actions forms a **Nash equilibrium** of a strategic game $\langle\{1,2\},(A_i),(U_i)\rangle$ if it has the property that neither player can profitably deviate, given the actions of the other players:

$$\text{(i)} \quad \neg\exists a_1' \in A_1 : \; U_1(a_1, a_2) < U_1(a_1', a_2)$$
$$\text{(ii)} \quad \neg\exists a_2' \in A_2 : \; U_2(a_1, a_2) < U_2(a_1, a_2')$$

For illustration, consider the following two games. In both games $N = \{1,2\}$, and $A_1 = \{a, b\}$, while $A_2 = \{c, d\}$. In both cases, it is optimal for player 1 (the row-player) to play a when player 2 (the column-player) plays c, and b when 2 plays d. The difference, however, is that in the first game player 2 strictly prefers c to d, while in the second game he strictly prefers d to c. It is easy to check that both games have exactly one Nash equilibrium, but that the equilibria are not the same: In the first game it is the profile (a, c), while in the second it is (b, d). The games and the Nash equilibria, the boxed payoffs, can be easily illustrated by the following matrices:

Game 1:

	c	d
a	**10,5**	0,0
b	0,6	5,5

Game 2:

	c	d
a	10,0	0,5
b	0,6	**5,10**

Profile (a, c) is a Nash equilibrium in the first game, because (i) if player 2 chooses c, it is best for player 1 to perform a, and (ii) if player 1 chooses a, it is best for player 2 to perform c. It is the *unique* Nash equilibrium of the game, because for all other strategy pairs at least one player has an incentive to deviate, given the choice of the other player.

In Dekker & van Rooy's (2000) reformulation of Bidirectional OT, the actions of speakers are thought of as the choice of expressions, and the actions of the hearers as the choice of interpretations. The Nash equilibria of such a game correspond with the form-meaning pairs that are *strongly optimal* in Blutner's sense. To account for his notion of *weak optimality*, and thus for Horn's division of pragmatic labor, Dekker & van Rooy (2000) make use of Nash equilibria in so-called *updated games* and show that the set of Nash equilibria in the fixed-point of such updated games correspond exactly with the weakly-optimal form-meaning pairs in Blutner's Bidirectional Optimality Theory.[5]

Although Dekker & van Rooy's game-theoretical interpretation of Bidirectional OT is appealing, the analysis is not completely satisfying. First of all, although the authors make use of the standard solution concept of a Nash equilibrium, this standard solution concept captures Blutner's crucial notion of weak optimality only when we consider the fixed point of the updated games. The notion of an 'updated game', however, is completely foreign to standard game theory, so it is not clear how far Horn's division of pragmatic labor really follows from general game-theoretical considerations. The reason why these updated games have to be considered is that the actions of the interpretation game participants are thought of as *concrete* forms and meanings. In particular, no justice is done to the fact that it depends on the situation the speaker is in which signal she sends and that the hearer chooses his interpretation only *after* he receives the form chosen by the speaker. It's assumed that in whatever situation the game is played, the equilibria will always be the same. There is indeed something to this when you want to account for *conventions* of language use, as (arguably) Horn's division of pragmatic labor does. But in the way things are implemented, it seems somewhat misleading to speak of speakers and hearers who play the game: all that is really done is to compare forms with meanings. This criticism carries over to Blutner's bidirectional OT. In the rest of the paper I want to take the role of speakers and hearers more seriously, consider the situations in which they are playing the game, and account for the fact that the game is a *sequential* one: interpretation comes

[5]The definition of an 'updated game' closely follows Jaeger's (2000) algorithm for computing optimal form-meaning pairs.

only after utterance.

3 Parikh on strategic communication

3.1 Description of the framework

Prashant Parikh (1991, 2000, 2001) gives a game-theoretic analysis of when communication is possible. He argues that speaker S communicates something to hearer H iff the discourse interaction can be described as what he calls a game of *partial* information with a *unique* solution. I will show in this section that by doing so Parikh in fact comes close to a general game-theoretical explanation of Horn's division of pragmatic labor.

Parikh wants to account for the fact that an, in principle, ambiguous or underspecified sentence like *Every ten minutes a man gets mugged in New York* typically gets interpreted as meaning that *some person or other* gets mugged every ten minutes, although it could, in principle, also get interpreted as meaning that *a particular man* gets mugged every ten minutes. Parikh intends to show under which circumstances only the first interpretation is part of a unique solution of the game played between speaker and hearer. In showing why this is so, Parikh makes use of alternative expressions, orderings of those expressions, and orderings of the possible meanings in a way that bears a close resemblance to Blutner's explanation of Horn's division of pragmatic labor.

In abstract, the argument roughly proceeds as follows: A speaker used an expression f that in principle could be interpreted in several ways. How f should, in fact, be interpreted depends on which actual situation the speaker is in, t or t'. With the sentence f, the speaker wants to communicate that she is in t, if she is in t, and t', if she is in t'. Although the speaker knows which situation she is in, the hearer does not, not even after f is uttered. The hearer may think now, for instance, that the speaker is in situation t with a probability of 0.8, and that she is in situation t' with a probability of 0.2. Moreover, suppose this is common knowledge. It might seem that for this reason the hearer should go for interpretation t, for this is the most likely interpretation. But, as with other coordination problems, things are not so simple: the hearer has to take into account the fact that the speaker used an expression expecting it to be interpreted by the hearer in the intended way, which once again depends on what the speaker could have used, and so on *ad infinitum*. Thus, just like Blutner, Parikh also assumes that for interpreting an expression,

we have also to take into account the *alternative expressions* that the speaker might have used. It is assumed that besides the underspecified form f, there are expressions f' and f'' that can each have one meaning only: f' can only mean t' and f'' only t. Parikh considers the act of interpreting form f as part of a larger game also involving the acts of interpreting the alternatives to f. He proposes that f can be interpreted as t only if this is the *unique solution* of the larger game.

Assuming that speaker and hearer want to communicate successfully, we can describe the situation as a *cooperative game* between speaker and hearer, where the speaker has private information about which state she is in that the hearer lacks, and where after receiving a form from the speaker, the hearer has to choose an action (interpret the form) that is either good for both or bad for both.

Although the action chosen by the hearer might depend on the action of the speaker, we might model the game as one in which they make their choices *simultaneously*. To do so, however, we have to assume that they choose **strategies** rather than concrete actions. A strategy consists of a rule that determines what a player will do in different circumstances. A speaker's strategy, S, is a function from situations to forms, i.e. an element of $[\{t, t'\} \rightarrow \{f, f', f''\}]$, and a hearer's strategy, H, is a function from forms to meanings/situations, i.e. an element of $[\{f, f', f''\} \rightarrow \{t, t'\}]$. In a table, this can be displayed as follows:

Speaker :

	t	t'
S_1	f	f'
S_2	f	f
S_3	f''	f
S_4	f''	f'

Hearer :

	f	f'	f''
H_1	t	t'	t
H_2	t'	t'	t

Thus, the search for equilibria is the search for an optimal combination of a speaker strategy and a hearer strategy. To be able to determine this we first have to know how the players *order* the profiles consisting of a sender-hearer strategy pair, $\langle S, H \rangle$. Parikh proposes to do this in terms of *expected utility*.

To get some intuition about why expected utilities might matter for games, let's go back for a moment to our abstract representation of the previous section. In the analysis of strategic games described there, we assumed that both players know the game they are playing. In particular, each player knows the payoffs for each player of every profile. This

suggests that each player has to know what the state of nature is. But this is not really necessary in order to use Nash equilibria as a solution concept. Suppose, for instance, that game 1 of the previous section is being played in state t, while game 2 is played in state t'. Suppose, moreover, that it is unknown to both players what the actual state is and thus what game is actually being played. However, it is commonly known that t is more likely to be played than t': $P(t) = 0.8$ and $P(t') = 0.2$. What counts in such a situation is not the actual payoffs in one particular game, but rather the *expected* payoffs. The expected payoff of profile (b, d) for player 2, $EU_2(b, d)$, for instance, is determined by $\sum_t P(t) \times U_2(t, b, d) = (0.8 \times 5) + (0.2 \times 10) = 6$. After calculating the expected utilities for both agents for all profiles, the game that is being played can be pictured as follows:

Cardinal:

	c	d
a	10,4	0,1
b	0,6	5,6

Ordinal:

	c		d
a	•	←	
	↑		↓
b		↔	•

So, to determine the equilibria of a game where the actual state is unknown to both players, we have to add a set of states, T, plus a probability distribution over these states. The payoffs of the profiles are thought of as expected utilities, or *lotteries*. Notice that not only are the payoffs of this game different from the original games, but also the expected plays: instead of one Nash equilibrium profile for each game we now have two of them in the combined one: (a, c) and (b, d).

Thus, just like for the game described above, we also need to know the probabilities and utilities involved in Parikh's game to determine expected utilities. Although the speaker knows which state she is in, Parikh implicitly assumes that this knowledge is not important to determine the utilities in terms of which the equilibria are calculated: the utility of underspecified form f depends on how the hearer is going to interpret it, and this, in turn, depends on the hearer's probability function over situations consistent with f's underspecified meaning, which is common knowledge. Thus, to determine expected utility, according to Parikh, only the hearer's probability function, P, is relevant. Expected utility is then determined in the usual way:

$$EU(S, H) \quad = \quad \sum_t P(t) \times U(t, S(t), H(S(t)))$$

Notice that in this definition I have followed Parikh, assuming that because communication is a game of coordination, speaker and hearer have the *same utility function*.[6] So, how should we define this function? Before we discuss Parikh's own proposal, it's instructive to first use a utility function that only distinguishes between successful and unsuccessful communication:[7]

$$\begin{aligned} U(t, S(t), H(S(t))) &= 1, \text{ if } H(S(t)) = t \\ &= 0 \text{ otherwise} \end{aligned}$$

Having fixed the probability and utility functions, we can calculate for each of the profiles its expected utility as a lottery over the utilities of the profiles in states t and t':

t:

t	H_1	H_2
S_1	1	0
S_2	1	0
S_3	1	1
S_4	1	1

t':

t'	H_1	H_2
S_1	1	1
S_2	0	1
S_3	0	1
S_4	1	1

partial:

	H_1	H_2
S_1	$\boxed{1}$	0.2
S_2	0.8	0.2
S_3	0.8	$\boxed{1}$
S_4	$\boxed{1}$	$\boxed{1}$

The idea now is that only the speaker-hearer strategy combinations that form a Nash equilibrium could be appropriate. We see that if we use a utility function that only cares about successful communication, the game described by Parikh has 4 Nash equilibria: $\langle S_1, H_1\rangle, \langle S_3, H_2\rangle, \langle S_4, H_1\rangle,$ and $\langle S_4, H_2\rangle$. Notice that we would get the same result for any other non-trivial probability distribution over the states: in all cases, the same 4 speaker-hearer strategy pairs would be Nash. Although in all 4 paired choices, communication would be successful, this game does not yet 'solve' the communication game: it remains unclear to the hearer how to interpret expression f. For successful communication the game must have a *unique solution*.

To do better than this, and to make use of the probability distribution over the states, Parikh proposes that the utility function is sensitive to the *complexity* of the expressions involved in the following way: successful communication is most important, but success with a simple expression (by using f) is preferred to success with a complex expression

[6] Parikh rightly notes that his analysis does not, and should not, depend on this assumption; it is made just to simplify matters.

[7] As we will see later, this is in fact the utility function used by Lewis (1969).

(by using f' or f''). Let us assume that the complexity of a form can be measured by a natural number and that $Compl(f) = 1$, while $Compl(f') = Compl(f'') = 2$. Making use of this complexity function, Parikh's utility function can be defined as follows:

$$\begin{aligned} U(t, S(t), H(S(t))) &= 1/Compl(S(t)), \text{ if } H(S(t)) = t \\ &= 0 \text{ otherwise} \end{aligned}$$

Now we can calculate for all the profiles their expected utilities as lotteries over the utilities of the profiles in states t and t' again, and see which profiles form Nash equilibria:

t:

t	H_1	H_2
S_1	1	0
S_2	1	0
S_3	0.5	0.5
S_4	0.5	0.5

t':

t'	H_1	H_2
S_1	0.5	0.5
S_2	0	1
S_3	0	1
S_4	0.5	0.5

partial:

	H_1	H_2
S_1	$\boxed{0.9}$	0.1
S_2	0.8	0.2
S_3	0.4	$\boxed{0.6}$
S_4	0.5	0.5

With this modified utility function, the game has two Nash equilibria: $\langle S_1, H_1 \rangle$ and $\langle S_3, H_2 \rangle$. Notice that according to the first one, $\langle S_1, H_1 \rangle$, the more probable state, or meaning, t, is expressed by the simple form f, while the less probable state, or meaning, t', is expressed by the complex form f'. Thus, $\langle S_1, H_1 \rangle$ might be called the *Horn strategy*. According to the other Nash equilibrium, however, the more probable meaning is expressed by a more complex form, while the less probable meaning is expressed by a lighter form, the anti-Horn strategy. Thus, only if speaker and hearer coordinate on the first Nash equilibrium we can give a game-theoretical explanation of Horn's division of pragmatic labor. But this means that in terms of Nash equilibria, we cannot yet account for Horn's division. Worse, because there is still more than 1 equilibrium left, we cannot yet even account for successful communication, because for that, according to Parikh, we have to show that the game has a *unique* solution. To solve the latter problem, Parikh proposes to refine the Nash equilibrium solution concept by taking only the *Pareto optimal* Nash equilibria into account. In our case this means that we select the Nash equilibrium which has the highest expected utility. Notice that in this way, Parikh also accounts for Horn's division: the solution of the game is $\langle S_1, H_1 \rangle$, the Horn-strategy pair according to which an (un)marked expression gets an (un)marked meaning, because that profile has a higher expected utility than $\langle S_3, H_2 \rangle$: 0.9 versus 0.6.

In the above example I have followed Parikh, giving concrete numbers to the probabilities of the states and the utilities of the form-meaning pairs in the states. But Parikh's analysis is not dependent on these particular numbers. All he needs for his analysis to go through is to make a distinction between successful communication with cheap (i.e. f) and with expensive (i.e. f' or f'') expressions. Let us assume that successful communication with simple expression f has benefit B, while with expressions f' or f'' there is an extra cost C involved, and the utility is thus $B-C$. To give the relevant tables, we will simplify things somewhat: note that in t, S_1 and S_2 give rise to the same behavior, i.e., f is sent, and that the same holds for S_3 and S_4, here f'' is sent. In t' something similar holds for, on the one hand, S_1 and S_4, and for S_2 and S_3 on the other. Thus, we might simplify the two speaker strategies in t, for example, to the use of the two forms f and f''. With this simplification, we can now give the following two qualitative tables:

t:

t	H_1	H_2
$S_1 = S_2 = f$	B	0
$S_3 = S_4 = f''$	$B-C$	$B-C$

t':

t'	H_1	H_2
$S_1 = S_4 = f'$	$B-C$	$B-C$
$S_2 = S_3 = f$	0	B

Denote the probability of state t by α and the one of state t' thus by $1-\alpha$. Now we can determine the expected utilities for each speaker-hearer combination in a qualitative way and see which combinations form Nash equilibria in the partial game. Making the natural assumption that $B > C > 0$, it is easy to see that $\langle S_1, H_1\rangle$ and $\langle S_3, H_2\rangle$ are again the only Nash equilibria and that they have the expected utilities $B-(1-\alpha)C$ and $B-\alpha C$, respectively. In case the probability of t is higher than the probability of t', i.e. $\alpha > (1-\alpha)$, strategy pair $\langle S_1, H_1\rangle$ is the unique Pareto optimal Nash equilibrium. Notice that in case $\alpha < (1-\alpha)$, i.e. when $\alpha < \frac{1}{2}$, strategy pair $\langle S_3, H_2\rangle$ would be the Pareto optimal one. And indeed, in that case the latter combination would be the one where the shortest expression gets the most likely interpretation. Thus, it depends on the probability distribution which strategy pair can be called the Horn strategy. Notice that when $\alpha = \frac{1}{2}$, both strategy pairs would be equally good, i.e. both would be Pareto optimal, and it would be unclear how to interpret the ambiguous expression f.

Notice that Parikh's use of a probability distribution over states and a complexity measure over forms gives rise to ordering relations over states and over forms that is used in a very similar way as Blutner uses his ordering relation in Bidirectional OT. Thus, both

Blutner and Parikh gave an analysis of how to interpret underspecified expressions in terms of orderings of meanings and forms. As already argued above, Parikh's analysis is not only more natural, but also more flexible: Blutner implicitly has to assume that communication is successful, while Parikh can explicitly measure the success of communication with the help of his utility functions. Despite these differences, the analyses give rise to the same interpretation mappings in the following sense: if t is more likely than t', $t > t'$ or $\alpha > (1 - \alpha)$, if for each state there are at least two forms that could express that state, and if f is the 'lightest' of those expressions, then both theories predict that t will be expressed by f and t' by a more complex expression.

Both analyses can be, or are, stated in game-theoretical terms. However, we saw that for Blutner's bi-OT a solution in terms of Nash equilibria did *not* generate *enough*, while for Parikh there were instead *too many* Nash equilibria. Dekker & van Rooy had to look at *updated games* to generate more acceptable form-meaning pairs, while Parikh had to select among the Nash equilibria to eliminate some possible solutions. Although this might seem puzzling, it's also not really very surprising: Blutner takes the actions to be the choice of *concrete* forms or meanings, while Parikh leaves the agents to choose between more general *strategies*. It's no wonder that the Nash equilibrium solution concept allows for many more acceptable form-meaning pairs in Parikh's more general framework.

3.2 Unsatisfying aspects of the framework

Although Parikh's game-theoretical account of Horn's division of pragmatic labor is more natural than the one of Dekker & van Rooy (2000) and uses more standard game-theoretical techniques, it is not completely satisfying either.

First, it shares some empirical problems with Blutner's (2000) Bidirectional OT: Suppose that f is a lighter expression than f', $f > f'$, and that f' can only mean t, but f can mean both. Suppose, moreover, that t is more salient, or stereotypical, than t', $t > t'$. In that case, Blutner's Bidirectional OT will predict that t' *cannot* be expressed: the form-meaning pair $\langle f, t \rangle$ will be weakly optimal, but there is no such pair for t'. Parikh predicts something similar. Suppose that higher salience of t versus t' is modeled via the following probability distribution: $P(t) = 0.8$ and $P(t') = 0.2$. This gives rise to the following tables:

	t	t'
S_1	f	f
S_2	f'	f

	f	f'
H_1	t	t
H_2	t'	t

t	H_1	H_2
S_1	1	0
S_2	0.5	0.5

t'	H_1	H_2
S_1	0	1
S_2	0	1

partial	H_1	H_2
S_1	0.8	0.2
S_2	0.4	0.6

In this case, Parikh's analysis predicts that there are 2 Nash equilibria: $\langle S_1, H_1 \rangle$ and $\langle S_2, H_2 \rangle$. Because the first one has a higher expected utility, it will be selected. But this prediction is not very satisfying: it means that just like Blutner, Parikh predicts that no sign will be interpreted as t'.[8] A concrete example which suggests that this prediction is wrong can be found in McCawley (1978). He notes that although pink is the stereotypical form of being pale red, the preexisting form 'pink' seems to have the effect that 'the dress is pale red' means that the dress is not prototypically pink, but somewhere between pink and pale red.

The second problem of Parikh's game theoretical account is that it uses an *unusual solution concept*: although the selection of the Pareto optimal Nash equilibrium seems natural, it is not one of the standard refinements of equilibria concepts that you find in the economic literature. In fact, there is a substantial amount of literature in economics discussing the question of how the Pareto optimal Nash equilibrium can be selected. This almost always involves extending the game by a round of communication (or *cheap talk*) before the actual game takes place. But a 'solution' of this kind seems very unnatural for our case, where the game itself is already about communication. As Parikh (1991) himself notes, such an approach suffers from the danger of an infinite regress.

Third, Parikh's suggestion that to account for successful communication and for Horn's division of pragmatic labor we must select the speaker-hearer strategy pair with the highest expected utility, in fact makes you wonder why he first introduces his quite involved game-theoretical setup in the first place. If we have to select the Pareto optimal Nash equilibrium, and if the payoffs for speaker and hearer are the same, things could be accounted for much simpler in terms of Shannon's (1948) Communication Theoretic principles of *optimal coding*.[9]

Finally, although the game that Parikh describes crucially involves *private information* – one individual has some information that the other lacks – I am not completely convinced

[8] It turns out that the analysis that I will propose in section 4 predicts rightly for this example.

[9] See van Rooy (to appear).

by the way he implements this: in a Nash equilibrium of the game, the hearer doesn't make full use of all the information he has. In particular, he is not making use of his knowledge of the strategy the speaker is using. Parikh does not analyze the game by using the standard techniques of solving such *games of private information*, also known as *Bayesian games*, in which this extra knowledge of the hearer is taken into account. In fact, Parikh (1991, p. 480) suggests that to analyze strategic inference in communication we have to think of new kinds of games and cannot use the tools developed to analyze games of private information. This is strange, because, by limiting ourselves to meanings/situations that underspecified f could denote, his analysis looks much like Lewis' (1969) well known analysis of *conventional meaning* in terms of the best studied games of private information: *signalling games*. I will show in the next section how far Parikh's analysis of strategic communication can be described in terms of games of private information, and in what sense his analysis of strategic communication is just like the strategic interactions involved in standard signalling games.

4 Signalling games

Quine (1936) challenged conventionalists' accounts of language to provide a satisfactory account of how the relevant conventions are set up and maintained that does not presuppose linguistic communication or competence. Lewis (1969) responded by explaining the semantic/conventional meaning of expressions in terms of equilibria of *signalling games*. In such games one player can send signals, or messages, to another player about the state the former player is in, but these messages have no pre-existing meaning; whatever meaning the messages acquire must emerge from the strategic interaction. Conventions, and conventional meanings, are then explained as stable Nash-equilibria.

Since Lewis introduced his signalling games to explain why and how conventional meanings can be associated with natural language expressions, these games have hardly been discussed within semantic and/or pragmatic analyses of natural language. In economics (and in biology, as we will see in section 5), however, generalizations of Lewis's signalling games have been studied extensively to throw light on, among others things, advertising and strategic pricing. In the next section I will describe a simple variant of signalling games as they are widely studied in economics, in terms of which we can also describe Parikh's situations of strategic interaction in communication.

4.1 Description of the framework

A signalling game is a two-player game with a *sender* and a *receiver*. This is a game of *private* information: The sender starts off knowing something that the receiver does not know. The sender knows the state t she is in but has no substantive payoff-relevant actions.[10] The receiver has a range of payoff-relevant actions to choose from but has no private information, and his prior beliefs concerning the state the sender is in are given by a probability distribution P over T; these prior beliefs are common knowledge. The sender, knowing t and trying to influence the action of the receiver, sends to the latter a signal of a certain form f drawn from some set F. The other player receives this signal, and then takes an action a drawn from a set A. This ends the game. Notice that the game is *sequential* in nature in the sense that the players don't move simultaneously: the action of the receiver might *depend* on the signal he received from the sender. The payoffs to the sender and the receiver are given by functions U_1 and U_2, respectively, which are elements of $[T \times F \times A \to R]$. For simplicity, we will assume here that T, F and A are all finite.

In the economics literature (e.g. Spence, 1973; Crawford & Sobel, 1982; Cho & Kreps, 1987) it is standardly assumed that the strategies of senders and receivers in signalling games are *probabilistic* in nature: the sender, for instance, is allowed to send different signals from within the same state, each with a certain probability, such that they add up to one.[11] For comparison with Parikh's analysis, however, I will simplify things, and assume that a *sender strategy*, S, is a *function* from states to signals (forms): $S \in [T \to F]$, and a *receiver strategy*, R, a function from signals to actions: $R \in [F \to A]$.

An *equilibrium* for a signalling game is described in terms of the strategies of both players. If the sender uses strategy S and the receiver strategy R, it is clear how to determine the utility of this profile for the sender, $U_1^*(t, S, R)$, in any state t:

[10] In game theory, it is standard to say that t is the *type* of the sender.

[11] We might be even more general and also take *noise* into account. This might be done by assuming that the relation between the forms that are sent and the forms that are received is also probabilistic in nature. Novak & Krakauer (1999) have recently argued that the formation of *words* and the need for *grammar* is natural and can be understood in a world where *mistakes* in implementation and comprehension are possible. This makes lots of sense, of course, from Shannon's (1948) communication-theoretic perspective: language rules increase the *redundancy* of the symbols used and this redundancy is useful in noisy communication channels, because it increases the *reliability* of the message sent.

$$U_1^*(t, S, R) \quad = \quad U_1(t, S(t), R(S(t)))$$

Due to his incomplete information, things are not as straightforward for the receiver. Because it might be that the sender using strategy S sends in different states the same signal, f, the receiver doesn't necessarily know the unique state relevant to determine his utilities. Therefore, he determines his utilities, or *expected* utilities, with respect to the *set* of states that he might be in after receiving message f. Let us define S_t to be the *information state* the receiver is in after the sender, using strategy S, sends her signal in state t, i.e. $S_t = \{t' : S(t') = S(t)\}$. With respect to this set, we can determine the (expected) utility of receiver strategy R in state t when the sender uses strategy S, $U_2^*(t, S, R)$:

$$U_2^*(t, S, R) \quad = \quad \sum_{t' \in S_t} P(t'/S_t) \times U_2(t', S(t'), R(S(t')))$$

A strategy profile $\langle S, R\rangle$ forms a *Nash equilibrium* iff neither the sender nor the receiver can do better by unilateral deviation. That is, $\langle S, R\rangle$ forms a Nash equilibrium iff for all $t \in T$ the following two conditions are obeyed:[12]

$$\begin{array}{ll} \text{(i)} & \neg\exists S' : U_1^*(t, S, R) < U_1^*(t, S', R) \\ \text{(ii)} & \neg\exists R' : U_2^*(t, S, R) < U_2^*(t, S, R') \end{array}$$

Let me stress again that the messages, or forms, used in these games have no pre-existing meaning. Meanings, so it was argued by Lewis (1969), could be associated with these messages, however, when, due to the sender and receiving strategies chosen in equilibrium, it is the case that the receiver acts differently (or appropriately, at least) when the sender is in different states. In that case we might say that the sender strategy S of the equilibrium pair $\langle S, R\rangle$ fixes meaning of expressions in the following way: for each state t,[13] the message $S(t)$ means t. But in order for this to be possible, it has to be the case that the game has an equilibrium $\langle S, R\rangle$ which indeed has the property that S sends

[12]Strictly speaking, this is not just a Nash equilibrium, but rather a *perfect Bayesian* equilibrium, the standard equilibrium concept for sequential, or *extensive form*, games with observable actions but *incomplete* information.

[13]Or a certain element of the partition of states

different messages in different states. Following standard terminology in economics, let us call $\langle S, R\rangle$ a *separating* equilibrium if it has this property. The following game, however, shows that not all signalling games have such an equilibrium.

4.2 Beer versus Quiche

Consider the signalling game due to Cho & Kreps (1987) where the receiver, Player 2, doesn't know the type of the sender, player 1: in state t player 1 is a surly fellow; in t' he is a wimp. Player 1 chooses whether to have beer or quiche for breakfast. Surly fellows prefer beer and wimps prefer quiche. After observing the breakfast chosen by player 1, player 2 decides whether to challenge agent 1 to a duel. Player 2 likes to fight wimps but fears fighting a surly fellow. Regardless of type, player 1 loses 1 unit of payoff if he has his less favorite breakfast, and he loses 2 units of payoff if he is challenged.

The following reasoning shows that this game has no separating equilibrium. Suppose there were a separating equilibrium, i.e., an equilibrium where a surly fellow and wimp have different breakfasts. Then the message sent by player 1 – the breakfast chosen – would have a 'meaning': it allows player 2 to infer player 1's type and make her decision whether to fight or not dependent on the message. Player 2 will choose to fight if she sees a wimp's breakfast and not to fight if she sees a surly fellow's breakfast. This means that either in state t or in state t' player 1 would be challenged to a duel. But this is always worse for player 1, even if he has the breakfast he prefers. This proves the nonexistence of a separating equilibrium.[14]

4.3 Lewis on conventional signalling

Above I have described signalling games as they are studied in economics. The games studied by Lewis (1969) are simpler in a number of respects. First, he assumes that the messages sent are *costless*. Formally this means that the utility functions are such that $U_i(t, f, a) = U_i(t, a)$ for both players i. In these circumstances, it turns out, the sender can only influence the receiver's decision of how to act if there is some common interest between the two.[15] This points to a second simplifying assumption made by Lewis: the

[14]The game does have two so-called 'pooling' equilibria, however. The pair $\langle S, R\rangle$ is called a *pooling* equilibrium, if there is a single signal f that the sender uses in all states.

[15]See Crawford & Sobel (1982).

interests of the players coincide: for every $t \in T$ and $a \in A$ it holds that $U_1(t, a) = U_2(t, a)$. For this reason we can work with one utility function U only. For ease of exposition, I will simplify matters even more and assume that the action of the receiver is just one of *interpretation*,[16] which means that the range of the receiver's strategy, A, equals the domain of the sender's strategy, T. Thus, I will assume that each sender strategy S is a function from states to forms: $S \in [T \to F]$, and each receiver strategy R a function from forms to states: $R \in [F \to T]$. The last special feature of Lewisian signalling games is that only successful communication counts for determining the utility of a sender-receiver strategy pair. Formally this means that for each t it holds that

$$\begin{aligned} U(t, S(t), R(S(t))) \quad &= \quad 1, \text{ if } R(S(t)) = t \\ &= \quad 0 \text{ otherwise} \end{aligned}$$

Such a game has several equilibria. A nice feature of Lewisian signalling games is that, if there are enough states and signals, equilibria are guaranteed to exist in which different signals are sent in different states which are interpreted appropriately. Such separating equilibria are called *signalling systems* by Lewis, and he proposes that these are the ones with which we associate linguistic meanings. These linguistic meanings can be called *conventional* if there are other competing signalling systems, or separating equilibria, that could have been chosen instead. In fact, a simple game with just two states, t and t', and two forms f and f', already has two separating equilibria: f is associated with t in the one, and with t' in the other.[17] Unfortunately, however, Lewisian games have many other equilibria besides these: they also always have a so-called *pooling* equilibrium in which the sender sends the same signal in all states, and a so-called *babbling* equilibrium in which the receiver ignores the utterance of the speaker and always 'responds' to the message sent by choosing the same action.[18] But if all these kinds of equilibria exist, in what sense are separating equilibria better than non-separating equilibria? One way of answering this question is in terms of *Pareto optimality*. Let us define the expected utility of a sender-receiver strategy $\langle S, R \rangle$ as before:

[16] In terms of Austin (1962): for meaning only *illocutionary* effects matter, not *perlocutionary* ones. See also Schiffer's (1972) criticism of Grice's (1957) analysis of non-natural meaning.

[17] There are many games with these states and forms, because in different games the receiver might have different probability distributions over the states.

[18] In distinction with separating equilibria, the resulting pooling and babbling equilibria crucially involve the probability function that represents the receiver's beliefs about the state the sender is in.

$$EU(S,R) \quad = \quad \sum_t P(t) \times U(t, S(t), R(S(t)))$$

It is easy to see that in case P assigns to at least 2 states a positive probability and there are at least two signals, the separating equilibria have a strictly higher expected utility than the non-separating equilibria. Thus, or so it seems, separating equilibria are chosen because they have the highest utility. Although there is something to this suggestion, we would like to give an explanation of it in terms of the *interaction* between sender and receiver, i.e., in terms of *game theory*. In the Lewisian signalling game, however, the expected utilities as such don't really play a role. Notice that the 'explanation' suggested above is the same as the one Parikh uses for selecting among several equilibria. Also he suggests that to account for successful communication and for Horn's division of pragmatic labor we must select the sender-receiver strategy pair which has the highest expected utility.

Even if we can give a game-theoretical account of why separating equilibria are better, we are not out of trouble. We have seen that a Lewisian game gives rise to several separating equilibria. But which one, then, will be chosen as the convention? Following Schelling (1960), Lewis suggests that this depends on which of those equilibria is most *salient*. But why would one separating equilibrium, or signalling system, be more salient than another? Perhaps, you might think, because one of them has the highest expected utility. Unfortunately, however, all separating equilibria of a Lewisian game are equal in this respect. How, then, can they be distinguished? Before we address this and the previous question, however, let us first show how Parikh could have used signalling games to analyze strategic communication.

4.4 Parikhian Signalling games

It should be clear already that Parikh's game-theoretical setup is very close to a signalling game: the speaker, *sender*, has private information about the state she is in that the hearer, *receiver*, lacks. The game is sequential in nature, and what counts are the *strategies* involved. Denoting the set of states, $\{t, t'\}$, by T and the set of forms, $\{f, f', f''\}$, by F, a sender strategy, S, is an element of $[T \rightarrow F]$ and a receiver strategy, R, an element of $[F \rightarrow T]$. So far, this is just like in Lewisian games. However, Parikh's games differ from Lewisian ones in a number of ways. First, Parikh wants to derive the meaning of only

one of three signals: he assumes that f' and f'' already have a unique fixed meaning. As we have seen in section 4, this means that we have to consider only 2 strategies of the receiver:

Sender :

	t	t'
S_1	f	f'
S_2	f	f
S_3	f''	f
S_4	f''	f'

Receiver :

	f	f'	f''
R_1	t	t'	t
R_2	t'	t'	t

Second, Parikh assumes that the signals used are directly relevant for the payoffs: his signalling game is not one of cheap talk meaning that the messages are not costless. As we have seen in section 3, the utility of a sender-receiver profile $\langle S, R\rangle$ in state t depends not only on whether communication is successful, $R(S(t)) = t$, but also on the complexity of signal $S(t)$:

$$\begin{aligned} U(t, S(t), R(S(t))) \quad &= \quad 1/Compl(S(t)), \text{ if } R(S(t)) = t \\ &= \quad 0 \text{ otherwise} \end{aligned}$$

The third difference concerns determining the (Nash) equilibria of the game. Let us look again at the games that, according to Parikh, are being played in the two states, and the resulting game. On the assumption that $P(t) = 0.8$ and $P(t') = 0.2$, Parikh assumes that these two games can be reduced to the one on the right hand side below, where the new payoffs are the expected utilities, and are assumed to be the same for both.

t:

t	R_1	R_2
S_1	1	0
S_2	1	0
S_3	0.5	0.5
S_4	0.5	0.5

t':

t'	R_1	R_2
S_1	0.5	0.5
S_2	0	1
S_3	0	1
S_4	0.5	0.5

partial:

	R_1	R_2
S_1	$\boxed{0.9}$	0.1
S_2	0.8	0.2
S_3	0.4	$\boxed{0.6}$
S_4	0.5	0.5

Parikh assumes that $\langle S, R\rangle$ is a Nash equilibrium of the whole game when it is a Nash equilibrium of the resulting game of 'partial' information on the right hand side.

When, however, we analyze communication as a signalling game, i.e. as a game with private information, the games played in t and t' are not really as described above, but

depend on what the receiver believes when a signal is sent to him. This latter belief, in turn, depends not only on the receiver's probability distribution over the states, but also on his knowledge of the sender's strategy in equilibrium (cf. the last criticism to Parikh in section 3.2). This means that the payoff functions U_1^* and U_2^* of sender and receiver, respectively, need not be the same, even though U_1 and U_2 are. For Parikh's example this only has an effect when the sender uses the same signal in both states, i.e., when she is using strategy S_2. Only in that case do the payoffs of the receiver depend on his prior belief about which state the sender is in. As a signalling game, the payoffs in each game look at follows:

t:

t	R_1	R_2
S_1	**1,1**	0,0
S_2	**1,0.8**	0,0.2
S_3	0.5,0.5	**0.5,0.5**
S_4	0.5,0.5	**0.5,0.5**

t':

t'	R_1	R_2
S_1	**0.5,0.5**	0.5,0.5
S_2	0,0.8	1,0.2
S_3	0,0	**1,1**
S_4	**0.5,0.5**	0.5,0.5

For $\langle S, R\rangle$ to be an equilibrium in a signalling game, it has to be a Nash equilibrium in all possible states, i.e., both in t and in t'. Although the solution concept used by Parikh is somewhat different from the one used in signalling games, it turns out that for the example under discussion, it doesn't really matter: in both cases the complete game has two equilibria: $\langle S_1, R_1\rangle$ and $\langle S_3, R_2\rangle$. In fact, given the setup of the games Parikh considers, there cannot be a difference in outcome.[19] I conclude that Parikh could just as well have analyzed his communicative situations in terms of standard signalling games.[20]

[19] A Nash equilibrium in Parikh's game can only disappear under a signalling game analysis if $\langle S_2, R_i\rangle$ has a higher utility than the equilibrium for player 1. But this is impossible because this player knows which state he is in. A Nash equilibrium in a signalling game but not in Parikh's corresponding partial game is impossible for the examples he considers, because the equilibria in the partial game are the only ones where (i) the strategy combinations have a positive payoff in both states, and (ii) the least complex expression f is involved.

[20] Actually, he *should* have analyzed things by using signalling games. Consider, again, the following problem for Parikh's analysis that we discussed earlier: with $f > f'$, $t > t'$, f' can only mean t, but f can mean both, and $P(t) = 0.8$ and $P(t') = 0.2$. This gives rise to the following tables.

	t	t'
S_1	f	f
S_2	f'	f

	f	f'
R_1	t	t
R_2	t'	t

t	R_1	R_2
S_1	1,0.8	0,0
S_2	0.5,0.5	0.5,0.5

t'	R_1	R_2
S_1	0,0.8	0,0
S_2	0.5,0	0.5,1

But wait! Parikh assumes that communication is successful only in case the game has a *unique* outcome, and doesn't he determine this outcome in terms of a notion, i.e. *expected utility*, that plays no role in signalling games? Yes, and no. True, you (almost) don't find expected utilities in the tables for t and t' above. But if we take over Parikh's assumption that the probabilities that the receiver assigns to the states are important for the payoffs of both the receiver *and* the *sender*, there is no reason why we could not calculate expected utilities in the same way as well. However, the signalling game reformulation of Parikh's framework strongly suggests that there is no good reason for doing so. But this means that Parikh's proposal to select the unique solution in terms of Pareto optimality is suspicious too. All that standard game theory can offer us when we describe the communication situation as a game between two rational players in a particular conversational situation is that one of the two equilibria should come out. To determine which one is not so much a matter of strategic inference in this particular situation, or so I would claim, but rather a matter of the players knowing a convention of language use that says that (un)marked expressions typically should be interpreted in (un)marked ways. But then, how could we explain this convention, how could it come about? To answer this question we will turn to evolutionary game theory.

5 Evolving Horn strategies

5.1 Evolutionary Game Theory

Until now we have not resolved the problem as to what equilibrium is likely to emerge. One way of resolving this problem is to introduce evolutionary or natural-selection considerations to game theory. But how could evolution explain game-theoretical equilibria? These equilibria are based on an analysis of *rational* choice, but what rational choices does an insect make? The idea behind evolutionary game theory (cf. Weibull, 1995) is that the players in a game are *not* taken to be the organisms under study, but rather the (genetic) *strategies* that both (i) determine the actual play of the organism, and (ii) replicate

We saw that Parikh's analysis predicts that $\langle S_1, R_1 \rangle$ should be chosen, which means that no sign will be interpreted as t'. According to the signalling game analysis, however, only $\langle S_2, R_2 \rangle$ will be a Nash equilibrium, and this is intuitively the right one. In this equilibrium f' means t and f means t'.

themselves. Payoffs are defined in terms of expected number of offspring, or replicants.[21]

Imagine a large uniform population of organisms who randomly encounter one another in pairwise interactions. In each match, each organism selects an action from the same set of possible modes of behavior. Each organism plays only once and leaves its offspring behind. The offspring of an organism playing a certain strategy depends on the strategy played by the organism with which it is matched. After many plays of the game, a strategy yielding a higher number of expected offspring will gradually come to be used by larger and larger fractions of the population. If the dynamic evolutionary process leads to a population all playing some single strategy such that mutants cannot invade it, then that strategy is evolutionarily stable. Maynard Smith & Price (1973) have characterized such evolutionarily stable strategies in the following way:

Strategy α is **evolutionarily stable** (an ESS) iff for all $\beta \in A$ it holds that

$$\begin{array}{ll} \text{(i)} & U_1(\alpha, \alpha) \geq U_1(\beta, \alpha), \text{ and} \\ \text{(ii)} & \text{if } U_1(\alpha, \alpha) = U_1(\beta, \alpha), \text{ then } U_1(\alpha, \beta) > U_1(\beta, \beta) \end{array}$$

Notice that the first condition guarantees that for strategy α to be evolutionarily stable, it has to be the case that profile (α, α) is a Nash equilibrium in the corresponding symmetric game between all strategies of A. However, it might be that (α, α) is a Nash equilibrium without α being evolutionarily stable: it might be that $U_1(\beta, \alpha) = U_1(\alpha, \alpha)$, but $U_1(\beta, \alpha) \leq U_1(\beta, \beta)$. This means that the standard equilibrium concept in evolutionary game theory is a *refinement* of its counterpart in standard game theory. Can this refinement be used to select between the several Nash equilibria in our signalling games discussed above?

5.2 Evolution in signalling games

The refinement can only be of some use if we think of sender-receiver strategy pairs in signalling games to be more the result of a kind of natural selection than the outcome of reasoned choice. But that doesn't seem to be a strange idea: a linguistic convention can be seen as a behavioral phenomenon that developed through the forces of evolution. Indeed,

[21] If A is the set of strategies, the expected number of offspring of an organism playing strategy α, $EU(\alpha)$, is $\sum_{\beta \in A} P(\beta) \times U_1(\alpha, \beta)$, where $P(\beta)$ is the proportion of organisms playing strategy β.

a linguistic convention can be thought of as a typical example of what Dawkins (1976) calls *memes*: cultural traits that are subject to natural selection. In contrast with genes, memes are not replicated – transmitted and sustained – through genetic inheritance, but through imitation, memory, and education. This has the result that cultural evolution can proceed much more rapidly than biological evolution. But linguistic conventions thought of as memes share a crucial feature with genes: if they do not serve the needs of the population, evolutionary forces will act to improve their functioning.[22]

> In order to understand how and why a language changes, the linguist must keep in mind two ever-present and antinomic factors: first, the requirements of communication, the need for the speaker to convey the message, and second, the principle of least effort, which makes him restrict his output of energy, both mental and physical, to the minimum compatible with achieving his ends. (Martinet, 1962, p. 139)

Thus, the idea is to think of sender-receiver strategy pairs as conventions that, if successful, can spread over a population through imitation or other kinds of adaptive behavior. A strategy pair is successful when (i) it accounts for successful communication, and (ii) it does so with small effort.

Let us first only consider the first condition for being successful, i.e. let us focus our attention on Lewisian utility functions. Look at signalling games with only two equally likely states: t and t'; two signals that the sender can use: f and f'; and two ways that the receiver can interpret signals: as a and as a', such that a corresponds with t, a' corresponds with t', and the utility function is Lewisian in the sense that only successful communication counts. Sender and receiver each have four strategies:

Sender :

	t	t'
S_1	f	f'
S_2	f	f
S_3	f'	f
S_4	f'	f'

Receiver :

	f	f'
R_1	a	a'
R_2	a'	a
R_3	a	a
R_4	a'	a'

If all individuals are of a single population, we can assume that they each adopt the role of sender and receiver half of the time. An individual's strategy must then consist

[22] See also Rubinstein's (2000) recent book.

of both a sender strategy and a receiver strategy. There are obviously 16 such strategy pairs. For each individual strategy α we can determine its expected payoff when it plays against strategy β. The following two tables show the payoffs of $\langle S_1, R_1\rangle$ and $\langle S_2, R_3\rangle$ in their sender (U_s) and receiver (U_r) role when they play against $\langle S_1, R_1\rangle$.

$\langle S_1, R_1\rangle$:

	U_s	U_r
t	1	1
t'	1	1

$\langle S_2, R_3\rangle$:

	U_s	U_r
t	1	1
t'	0	0

Notice that the expected payoff of strategy $\langle S_1, R_1\rangle$ playing against itself is $(\frac{1}{2} \times (\sum_t P(t) \times U_s(t))) + (\frac{1}{2} \times (\sum_t P(t) \times U_r(t)))$, which is $(\frac{1}{2} \times ((0.8 \times 1) + (0.2 \times 1))) + (\frac{1}{2} \times ((0.8 \times 1) + (0.2 \times 1))) = 1$. This is strictly higher than the expected payoff of strategy $\langle S_2, R_3\rangle$ playing against $\langle S_1, R_1\rangle$, which is $(\frac{1}{2} \times ((0.8 \times 1) + (0.2 \times 0))) + (\frac{1}{2} \times ((0.8 \times 1) + (0.2 \times 0))) = 0.8$. In fact, there is no strategy that plays as good as or better against strategy $\langle S_1, R_1\rangle$ than this strategy itself. This means that $\langle S_1, R_1\rangle$ is evolutionarily stable. Doing all the calculations for all the strategies, it is easy to see that there are only two individual strategies that are evolutionarily stable: $\langle S_1, R_1\rangle$ and $\langle S_3, R_2\rangle$. These two individual strategies are exactly the two *separating Nash equilibria* in the corresponding Lewisian signalling game. As shown by Waerneryd (1993), something more general holds: For any sender-receiver game of the kind introduced above, with the same number of signals as states and actions, a strategy is evolutionarily stable if and only if it is a separating Nash equilibrium.[23] In this way Waerneryd has given us a beautiful and purely game-theoretical

[23] You might wonder: does this result not rule out the possibility of *synonymy* and *ambiguity*? No it does not. Suppose that we have one more form than we have meanings. In that case, we can assume that one meaning can be expressed by two different forms and we thus have *synonymy*. The optimal language with synonymy would be one where the speaker's strategy is probabilistic in nature and assigns to one meaning two different forms with a positive probability. The hearer's strategy, however, is not probabilistic: it assigns to both forms the same meaning. This speaker-hearer strategy pair is also evolutionarily stable. As for *ambiguity*, we have to consider Parikh's game again. Here a separating equilibrium could be reached with just the two more complex expressions f' and f''. However, enriching the language with the potentially ambiguous f is successful: in each communicative situation we still have a separating equilibrium, but now one with on average a higher utility. Thus, our analysis suggests that ambiguity is possible, and profitable, but only with 'cheap' expressions.

explanation of why separating Nash equilibria should evolve.[24,25]

Notice, however, that even for the simple signalling game under discussion, there are already two stable strategies. But only one of the two will evolve. Lewis (1969) suggested that this will be the *salient* one. But there is no reason for one to be more salient than the other. Can't we give a purely game-theoretical explanation of selection among the equilibria? Skyrms (1996) shows that we can if we also take into account the *dynamic process* by which such stable states can be reached.

This so-called *replicator dynamics* (Taylor & Jonker, 1978) shows us that the strategy which dominates the population at a fixed point of the dynamic process, the evolutionarily stable strategy,[26] depends on the proportion of the different strategies in the population in the *initial state*. Different initial states might give rise to different fixed points in the replicator dynamics, and thus to the evolution of different evolutionarily stable strategies. We have seen before that it is *necessary* that a separating Nash equilibrium will evolve in our signalling games. Now we see that the particular one that will evolve is a matter of *chance*. Skyrms (1996) concludes that if the evolution of linguistic conventions proceeds as in replicator dynamics, there is no need to make use of the psychological notion of *salience* to explain selection of equilibria.

5.3 The evolution of alarm calls

We have seen above that to account for Horn's division of pragmatic labor, Parikh could not rely only on the standard game-theoretic solution concept of a Nash equilibrium, but had also to make use of the notion of Pareto optimality. Parikh discussed a signalling game with 2 states, 3 forms (or signals), and 2 meanings (or actions). Two sender-receiver pairs constituted a Nash equilibrium, but one of the two had a higher expected utility than the other. Can we give a natural game-theoretic explanation of why we, in fact, only see

[24]For simplicity I have considered only pure strategies. Skyrms (1996) shows that if we also consider *mixed* strategies, we need to take into account the replicator dynamics to explain why only separating Nash equilibria will evolve.

[25]As stressed by Skyrms (1996), Waerneryd shows even more: Lewis's (1969) requirement of common knowledge is not needed to explain why a linguistic signalling convention (i.e., a stable strategy in an evolutionary signalling game) can be sustained.

[26]Where the fixed points are really the *asymptotically stable* points, and where we consider real or 'reduced' 2×2 games. For explanation of these notions and a more detailed characterization, see Weibull (1995).

the former equilibrium – which respects Horn's division – but not the latter? The notion of an evolutionarily stable strategy is more fine-grained than that of a Nash equilibrium, so perhaps we can find the solution here, especially when we also take *effort* (costs) into account.

Noble (2000) has recently given an evolutionary account of the emergence of a signalling system that seems to correspond closely to what we are after. Noble wants to explain why animals send signals if they are in certain situations (when there is food, or danger) and not in others. According to Noble's signalling game, we have 2 states, t and t', the sender sends either signal f or f', and the receiver chooses either a or a', where only the latter is appropriate (useful) for both in t'. These signalling games come with 4 sender strategies and 4 receiver strategies, giving us 16 combined sender-receiver strategies.

So far this is exactly like the signalling games we have been discussing above. However, he makes some extra assumptions: (i) sending signal f is cost-free, but sending f' is not; (ii) taking action a is cost-free, but taking action a' is not; and (iii) the sender is ambivalent about the receiver's response in state t.

Noble's purpose in describing this signalling game was to determine under which conditions the *honest* strategy $\langle S_1, R_1 \rangle$ is the only one that is evolutionarily stable. Because this is the strategy that reminds us of the Horn strategy, i.e., the strategy that implements Horn's division of pragmatic labor, it seems that Noble's characterization of the conditions is also highly relevant for us. He calculates that this is the case iff the benefits of successful communication are higher than the costs involved.

Can we conclude from Noble's discussion that this characterization also accounts for Horn's division of pragmatic labor? Unfortunately, we cannot, because of one crucial difference between our models and the ones described by Noble. In the latter the sender is supposed to be *ambivalent* about the receiver's response in state t, but this doesn't make much sense in normal communicative situations. In normal communicative situations both the sender and the receiver prefer that the receiver perform the one appropriate action a in state t, i.e., interprets the signal as 'meaning' t.[27]

We might try to fix that by removing the assumption that the sender is ambivalent about state t, so that now both the sender and the receiver get positive payoffs when the

[27] For a similar reason, Spence's (1973) solution of signalling games involving costs is of no use for our analysis either: Spence crucially assumes that one of the receiver's actions is preferred by the signaller, in whatever state she is in.

receiver performs action a. Unfortunately, however, when we make this move, strategy $\langle S_1, R_1 \rangle$ is no longer the only one that is evolutionarily stable: this is also the case for sender-receiver strategy $\langle S_3, R_2 \rangle$. This should not really surprise you, for we have seen above that both strategy pairs are separating Nash equilibria. We are back to where we started.

Our first try to characterize Horn's division of pragmatic labor in evolutionary terms failed. The notion of an ESS – even when we took costs into account – did not bring us what we had hoped for. There is actually a very general reason why it *could* not have worked. Think again of $\langle S_1, R_1 \rangle$ and $\langle S_3, R_2 \rangle$ as sender-receiver strategy pairs in the non-evolutionary setting. By determining the utilities as proposed by Parikh, both are (separating) Nash equilibria. But, in fact, they are even stronger than that: they are *strict* Nash equilibria. Profile (α, β) is a strict Nash equilibrium if there is no α' such that $U_1(\alpha, \beta) \leq U_1(\alpha', \beta)$ and no β' such that $U_1(\alpha, \beta) \leq U_1(\alpha, \beta')$. We have seen above that it's a *necessary* condition for α to be evolutionarily stable that (α, α) is a Nash equilibrium in a symmetric game between all strategies. It is easy to see that it's also a *sufficient* condition for α to be evolutionarily stable that (α, α) is a *strict* Nash equilibrium. But then it follows that not only $\langle S_1, R_1 \rangle$ but also $\langle S_3, R_2 \rangle$ *must* be evolutionarily stable in the sense of being an ESS.[28]

Where does this leave us? We tried to get rid of the unwanted Nash equilibrium profile $\langle S_3, R_2 \rangle$ by using evolutionary game theory, but we have just seen that there is no hope of doing so by using its standard solution concept.

But perhaps this just means that we were too demanding. Perhaps we should permit the anti-Horn strategy to be evolutionarily stable, but, as suggested by one of the reviewers, explain its non-occurrence in a more circumstantial way. As we saw in section 5.2, although each of the evolutionarily stable strategies *can* evolve, which one of them actually *will* evolve depends on the *initial distribution* of the individual strategies. At present, it is unclear to me how far such an analysis would get us. In the next section I will discuss other possible ways to go that explain the emergence of the Horn strategy in a more analytic way.

[28] In the second chapter of Rubinstein (2000), several alternative evolutionarily solution concepts are discussed, some of them involving complexity considerations. However, in case (α, α) is a strict Nash equilibrium, it will also be evolutionarily stable according to all these alternative solution concepts.

5.4 The evolution of Horn strategies: correlation and mutation

Think of a situation where individuals have the choice between the two sender-receiver strategies that were separating Nash equilibria in the signalling game suggested by Parikh: $\langle S_1, R_1 \rangle$ and $\langle S_3, R_2 \rangle$. In the evolutionary setting we then have a symmetric 2×2 coordination game with the following payoff matrix:

	$\langle S_1, R_1 \rangle$	$\langle S_3, R_2 \rangle$
$\langle S_1, R_1 \rangle$	0.9	0.25
$\langle S_3, R_2 \rangle$	0.25	0.6

Notice that when both are playing the same sender-receiver strategy, they get exactly the payoff that Parikh calculated as its expected utility in his game of partial information. Thus, although these expected utilities didn't really play a role in a standard signalling analysis, they are crucial in an evolutionary setting.

Our problem was that although the second Nash equilibrium is Pareto-dominated by the first, it is still evolutionarily stable according to the standard stability concept and its corresponding dynamic process. When we give up some assumptions behind this replicator dynamics – *random pairing* and *determinism* –, it turns out that we can explain why we actually end up with only the Pareto optimal equilibrium. I will first discuss the consequences of assuming the possibility of positive *correlation* and then discuss an evolutionarily process where *mutations* are allowed.

It is well known in biology that vervet monkeys, *Cercopithecus aethiops*, use vocal alarm signals to warn their fellow troop members of at least three quite distinct kinds of predators. Different alarms are given for different kinds of predators, and at different alarm calls their fellow troop members respond in different ways. This looks much like communication and we even see a separating Nash equilibrium at work here: a Lewisian signalling system. As we have seen in section 5.2, such signalling systems can evolve if payoffs of sender and receiver are equal. This, however, doesn't seem to be the case here. In a community of vervet monkeys it is profitable to be a receiver: you are alerted for predators. But how could it be profitable for a vervet monkey to *send* an alarm call? The monkey already knows about the predator, and giving the alarm call not only costs energy, but might even attract the attention of the predator. In fact, it turns out that if utility

is just measured in terms of fitness, the strategy $\langle S_1, R_1 \rangle$ in the signalling game played by vervet monkeys is *not* evolutionarily stable because it can be invaded by the *free rider* strategy $\langle S_2, R_1 \rangle$ that does not send, but reacts appropriately. Why don't we see free riders in vervet monkeys? The reason is, or so it is suggested by Skyrms (1996), that 'honest' monkeys *don't pair at random* with free riders, although this random pairing is implicitly assumed in the replicator dynamics behind the ESS solution concept. Skyrms shows that if we assume that there exists enough of a positive correlation between pairings of similar sender-receiver strategies, free riders can be driven to extinction by honest senders. In fact, with enough positive correlation, 'honest' strategy $\langle S_1, R_1 \rangle$ cannot be invaded by free rider $\langle S_2, R_1 \rangle$ and is evolutionarily stable again, but now in a more general sense. Skyrms (1994) defines a more general stability concept for evolutionary settings, *adaptive ratifiability*,[29] and shows that if a strategy is adaptively ratifiable, then it is an attractive equilibrium in the more general replicator dynamics where random pairing is no longer assumed and correlation is possible.

How does this help us to account for the linguistic convention that (un)marked forms are typically associated with (un)marked meanings? Well, note first that positive correlation is one possibility to explain a tendency towards Pareto-efficiency in games.[30] Let us call strategy α *strictly efficient* if in interaction with itself it has a higher utility than any other strategy β in selfinteraction: $U(\alpha, \alpha) > U(\beta, \beta)$. Then Skyrms (1994) shows that when correlation is (nearly) perfect, the strictly efficient strategy is (the unique) equilibrium of the replicator dynamics.[31] Second, it seems natural to assume that for linguistic communication, positive correlation is the rule rather than the exception: we prefer and tend to communicate with people that use the same linguistic conventions as we do, otherwise communication would fail in lots of circumstances. We might also think of language *learning* and education as giving rise to correlation. In fact, there might even be a deeper connection between certain 'educational' learning models and evolution with positive correlation. For our purposes this means that, others things being equal, when people tend

[29] The name, and idea behind it, comes from the optimality concept in Jeffrey's (1983) *evidential* decision theory where it is assumed that the state of nature depends probabilistically on the act (modeled as a proposition) chosen.

[30] The idea that positive correlation, or *viscosity*, is favorable to the evolution of altruism goes back to Hamilton (1964) and is popularized by Dawkins (1976) and Axelrod & Hamilton (1981).

[31] Yes, it has the consequence that even players of the prisoners' dilemma game will reach the efficient (but non-Nash) outcome, if they correlate their behavior enough.

to speak only with others who use the same linguistic conventions, it is predicted that – slowly but surely – only strictly efficient linguistic conventions will evolve. In particular, it means that communities that use the anti-Horn strategy will die out leaving only communities that use the Horn strategy. Thus, we can give an evolutionary explanation of Horn's division of pragmatic labor.

As it turns out, there is another purely analytic way to explain why the Horn strategy has evolved.[32] Standard replicator dynamics assumes that evolution is a deterministic process. Kandori, Mailath & Rob (1993) study an indeterministic variant of this process in which a stochastic component, analogous to *mutations* in biological models, plays a role. Such a process might have various equilibria – population states which have no tendency to change – but the added random element has the effect that populations can escape from such equilibria and the system flips to another stable point. However, some of these stable points are more robust than others: the most robust equilibrium is the one that requires the highest number of mutations to escape from. In case the mutation rate is small, the system spends most of the time at the 'good' equilibrium, with probability 1 in the long run. In fact, Kandori, Mailath & Rob's *long run equilibrium* concept coincides with the so-called *risk dominant* equilibrium which in many games – including the one we are discussing – coincides with the Pareto optimal one.

If we think of linguistic innovation as the random element in the dynamic model sketched above, we can use the results to explain the emergence of Horn strategies. If the rate of innovation is small relative to the population size, Horn strategies will come to dominate anti-Horn strategies in the long run.[33,34] Thus, not only correlation, but also mutation can explain the evolution of Horn's division of pragmatic labor.

[32] I am indebted to Taub & Hoerner (ms) for pointing this out to me. They argue in favor of this alternative solution partly on the grounds that my exclusive use of Skyrms' notion of adaptive ratifiability in an earlier version of this paper is not a refinement of standard equilibrium concepts.

[33] See Taub & Hoerner (ms) for more detail.

[34] Some AI students (Erik Borra, Tom Lentz, Arnold Obdeijn, Jasper Vijlings and Reinier Zevenhuijzen) at the university of Amsterdam simulated the evolutionary process by making use of a genetic algorithm. Starting out from a number of initial states, including one where all players play the anti-Horn strategy, they found a clear bias for the Horn strategy. Just like Kandori, Mailath, & Rob (1993), they made crucial use of mutations in their genetic algorithm. At present, however, it is not clear to me whether the similarity in results and background ideas are not based on a more analytic similarity. See `http://home.student.uva.nl/reinier.zevenhuijzen/signalling/` for the experiment they conducted and play with it yourself.

6 Conclusion and Outlook

In this paper I have discussed how Blutner's bidirectional optimality explanation of Horn's division of pragmatic labor can be accounted for within Parikh's game-theoretical analysis of successful communication. Moreover, I have shown how the latter analysis relates to Lewisian signalling games and proposed that the observation that (un)marked expressions typically get an (un)marked meaning is conventional rather than conversational in nature and should, in the end, be accounted for in evolutionary terms. I also suggested some particular ways in which Horn's strategy might have evolved, but I have not given evidence for why one solution would be more plausible than the other.

I didn't discuss many concrete phenomena and limited myself to the analysis of some simple examples discussed in the introductory part of this paper. However, I believe that the techniques discussed in this paper have many more potentially interesting applications.

A first possible application of the present framework is to give a game-theoretical reformulation of Benz's context-sensitive bidirectional OT: Blutner assumes that there exists a single ordering '>' between form-meaning pairs that is valid for both speaker and hearer. Benz (ms) studies situations where this assumption no longer holds and extends bi-OT by taking context into account. However, we noted that in contrast to bidirectional OT, game theoretical analyses of communication assume that the game is situated. This suggests that we can simply borrow ideas from standard game theory to deal with Benz' situations as well.

Another application would be to use the evolutionary framework to derive other pragmatic rules of language use than just Horn's division of labor. In fact, there already exists a substantial literature in both economics and biology explaining the evolution of 'honest' sender-receiver strategy pairs (see also sections 5.3 and 5.4 of this paper) and Asher, Sher & Williams (2001) propose to give a game theoretical foundation for Grice's maxim of Quality.

In this paper I took for granted a particular interpretation of the probability and utility functions involved. To account for the full range of phenomena subject to Horn's division of pragmatic labor, or even to extend that division, we could, or should, interpret the probabilities and utilities involved in a more flexible way.

As for probabilities, I simply assumed that they measure the stereotypicality of the states. A straightforward extension would be to interpret the probabilities as governing

salience orderings to account for pronoun resolution. Recent work of Zeevat & Jaeger (2002) suggest a somewhat more complex extension: the prior probability of the contents can be derived from the corpus-based statistical relations between forms and meanings. An exciting interpretation of the probability function is already proposed by Taub & Hoerner (ms): they assume that the probabilities involved measure the likelihood of the *frames*, or *scripts*, used to interpret potentially ambiguous expressions. In this way they use game theory to extend standard cognitive semantics.

In this paper I have assumed that evolutionary game theory is useful to account for the emergence of certain conventions. This evolutionary analysis is closely related with approaches like that of Novak & Krakauer (1999) that seek to explain the emergence of grammar. However, it might also be useful to account for the further development of grammars, i.e. the process of *grammaticalization* and *language change* (e.g. Croft, 2000). Throughout the paper I have made the simplifying assumption that the optimal (use of) language depends only on the complexity of the expressions used. However, the game-theoretical framework, as such, is much more general: there is no reason to assume that the utility function cannot (also) depend on other factors. A simple modification would be to also take into account the fact that ambiguous expressions are normally more difficult to interpret than unambiguous ones. In more exiting applications we could perhaps also take notions such as *politeness* of expression and *relevance* of information into account.[35]

References

[1] Asher, N., I. Sher and M. WIlliams (2001), 'Game theoretical foundations for Griean constraints', In R van Rooy & M. Stokhof (eds.), *Proceedings of the Thirteenth Amsterdam Colloquium*, ILLC, Amsterdam.

[2] Atlas, J. and S. Levinson (1981), 'It-Clefts, Informativeness and Logical Form', In: P. Cole (ed.), *Radical Pragmatics*, New York, AP.

[3] Austin, J. L. (1962), *How to do things with words*, Oxford University Press.

[35] The cost of ambiguity is taken into account by Taub & Hoerner (ms). If we also consider politeness and relevance, it could be that the evolutionary trajectory is not as uni-directional as the proposals discussed in section 5 seeked to account for.

[4] Axelrod, R and W. Hamilton (1981), 'The evolution of cooperation', *Science*, **411**: 1390-6.

[5] Benz, A. (ms), 'Bidirectional Optimality Theory with Context-Sensitive Constraints', Humbold Universitaet Berlin. Available from ROA 465-0901.

[6] Blutner, R. (2000), 'Some aspects of optimality in natural language interpretation', *Journal of Semantics*, **17**: 189-216.

[7] Cho, I. and D. Kreps (1987), 'Signalling games and stable equilibria', *The Quarterly Journal of Economics*, **102**: 179-221.

[8] Crawford, V.P. and J. Sobel (1982), 'Strategic information transmission', *Econometrica*, **50**: 1431-1451.

[9] Croft, W. (2000), *Explaining Language Change: An Evolutionary Approach*, Longman Linguistics Library, Harlow.

[10] Dawkins, R. (1976), *The Selfish Gene*, Oxford: Oxford University Press.

[11] Dekker, P. and R. van Rooy (2000), 'Bi-Directional Optimality Theory: An application of Game Theory', *Journal of Semantics*, **17**: 217-242.

[12] Grice, P. (1957), 'Meaning', *Philosophical Review*, **66**: 377-88.

[13] Grice, P. (1975), 'Logic and conversation', In: P. Cole & J.L. Morgan (eds.), *Syntax and Semantics, 3: Speech Acts*, New York: AP.

[14] Hamilton, W. D. (1964), 'The genetic evolution of social behaviour', *Journal of Theoretical Biology*, **7**: 1-52.

[15] Hendriks, P. and H. de Hoop (2001), 'Optimality Theoretic Semantics', *Linguistics and Philosophy*, **24**: 1-32.

[16] Horn, L. (1984), 'Towards a new taxonomy of pragmatic inference: Q-based and R-based implicature'. In: D. Schiffrin (ed.), *Meaning, Form, and Use in Context: Linguistic Applications*, GURT84, 11-42, Washington; Georgetown University Press.

[17] Jaeger, G. (2000), 'Some notes on the formal properties of bidirectional optimality theory', In: R. Blutner & G. Jaeger (eds.), *Studies in Optimality Theory*, Linguistics in Potsdam, volume 8, pp. 41-63.

[18] Jeffrey, R. (1983), *The logic of Decision*, 2d revised ed., Chicago: Chicago University Press.

[19] Kandori, M., G.J. Mailath and R. Rob (1993), 'Learning, mutation, and long run equilibria in games', *Econometrica*. **61**: 29-56.

[20] Kreps, D. and J. Sobel (1994), 'Signalling', In: R.J. Aumann & S. Hart (eds.), *Handbook of Game Theory, Vol. 2*, Elsevier.

[21] Levinson, S. *Presumptive Meanings. The Theory of Generalized Conversational Implicatures*, MIT Press: Cambridge, Massachusetts.

[22] Lewis, D. (1969), *Convention*, Cambridge: Harvard University Press.

[23] McCawley, J.D. (1978), 'Conversational Implicatures and the Lexicon', in P. Cole (ed.), *Syntax and Semantics, vol 9: Pragmatics*, Academic Press, New York.

[24] Martinet, A. (1962), *A functional view of language*, Oxford: Clarendon Press.

[25] Maynard Smith, J. and G.R. Price (1973), 'The logic of animal conflict', *Nature*, **146**: 15-18.

[26] Noble, J. (2000), 'Cooperation, competition and the evolution of prelinguistic communication', In: C. Knight et al (eds.), *The Emergence of Language*, pp. 40-61, Cambridge University Press.

[27] Novak, M. and D. Krakauer (1999), 'The evolution of language', *Proc. Natl. Acad. Sci. USA*, **96**: 8028-8033.

[28] Parikh, P. (1991), 'Communication and strategic inference', *Linguistics and Philosophy*, **14**: 473-513.

[29] Parikh, P. (2000), 'Communication, meaning, and interpretation', *Linguistics and Philosophy*, **23**: 185-212.

[30] Parikh, P. (2001), *The use of Language*, CSLI Publications, Stanford, California.

[31] Quine, W.V.O. (1936), 'Truth by convention', In: O. H. Lee (ed.), *Philosopohical Essays for A. N. Whitehead*, New York: Longmans.

[32] Rooy, R, van (to appear), 'Conversational implicatures and Communication Theory', In. J. van Kuppevelt & R. Smith (eds.), *Current and New Directions in Discourse and Dialogue*, Dordrecht (Kluwer).

[33] Rubinstein, A. (2000), *Economics and Language*, Cambridge University Press.

[34] Schelling, T. (1960), *The Strategy of Conflict*, New York: Oxford University Press.

[35] Schiffer, S. (1972), *Meaning*, Clarendon Press, Oxford.

[36] Shannon, C. (1948), 'The Mathematical Theory of Communication', *Bell System Technical Journal*, **27**: 379-423 and 623-656.

[37] Skyrms, B. (1994), 'Darwin meets *the logic of decision*: Correlation in evolutionary game theory', *Philosophy of Science*, **61**: 503-528.

[38] Skyrms, B. (1996), *Evolution of the Social Contract*, Cambridge University Press.

[39] Spence, M. (1973), 'Job market signalling', *Quarterly Journal of Economics*, **87**: 355-74.

[40] Taub, S. and A. Hoerner (ms), Game theory and semantic frames, manuscript.

[41] Taylor, P. and L. Jonker (1978), 'Evolutionary stable strategies and game dynamics', *Mathematical Biosciences*, **40**: 145-56.

[42] Waerneryd, K. (1993), 'Cheap talk, coordination, and evolutionary stability', *Games and Economic Behavior*, **5**: 532-546.

[43] Weibull, J. W. (1995), *Evolutionary Game Theory*, MIT Press, Cambridge.

[44] Zeevat, H. and G. Jaeger (2002), 'A reinterpretation of syntactic alignment', in de Jongh et al. (eds.), *Proceedings of the Fourth Tbilisi Symposium on Language, Logic and Computation*. Tbilisi/Amsterdam.

[45] Zipf, G. (1949), *Human behavior and the principle of least effort*, Cambridge: Addison-wesley.