Games and Quantity Implicatures*

Robert van Rooij

Abstract

In this paper we seek to account for scalar implicatures and Horn's division of pragmatic labor in game-theoretical terms by making use mainly of refinements of the standard solution concept of signaling games. Scalar implicatures are accounted for in terms of Farrell's (1993) notion of a 'Neologism-Proof' equilibrium together with Grice's maxim of Quality. Horn's division of pragmatic labor is accounted for in terms of Cho & Kreps' (1987) notion of 'equilibrium domination' and their 'Intuitive Criterion'.

Keywords: Conversational Implicatures, Game Theory, Horn's division, Intuitive Criterion, Neologism Proofness, Pragmatics

1 Implicatures and Grice's maxim of Quantity

Perhaps the most important notion in linguistic pragmatics is Grice's (1967) notion of *conversational implicature*. It is based on the insight that by means of general principles of rational communication we can communicate more with the *use* of a sentence than the *conventional meaning* associated with it. What is communicated depends not only on syntactic and semantic rules, but also on facts about the utterance situation, the linguistic context, and the goals and preferences of the interlocutors of the conversation. These implicatures are based on Grice's *cooperative principle*: the assumption that speakers are maximally efficient rational cooperative language users. Grice comes up with a list of four rules of thumb – the maxims of *Quality*, *Quantity*, *Relevance*, and *Manner* – that specify what participants have to do in order to satisfy this principle. They should speak sincerely, relevantly, and clearly, and should provide sufficient information.

Over the years many phenomena have been explained in terms of the Gricean maxims of conversation. Horn (1972) and especially Gazdar (1979)

*I would like to thank Michael Franke, Kris de Jaegher, Tikitu de Jager, Brian Skyrms, and especially the anonymous reviewer for this journal for remarks and discussion.

proposed to formalize Grice's suggestions in order to turn informal pragmatics into a predictive theory. They concentrated on Grice's maxim of Quantity, and especially on its first submaxim.

- **Quantity**

  1. Make your contribution as informative as is required (for the current purposes of the exchange).
  2. Do not make your contribution more informative than is required.

The first submaxim is said to induce inferences from the use of one expression to the assumption that the speaker did not intend to communicate a contrasting, and informationally stronger, one. It is used to motivate the inferences from 'John ate *some* of the cookies' and 'John has two children' to 'It is *not* the case that John ate *all* of the cookies' and 'John has *exactly* two children', respectively. Atlas & Levinson (1981), Horn (1984) and Blutner (2000) tried to formally account not only for Quantity_1 implicatures, but also for implicatures that appeal to Grice's second Quantity maxim and the maxims of Relation and Manner. The most important type of implicature that they claim to be able to account for by taking also the second submaxim of Quantity into account is *Horn's division of pragmatic labor*, according to which an (un)marked expression (morphologically complex and less lexicalized) typically gets an (un)marked meaning. In this paper we want to investigate how Quantity_1 implicatures and Horn's division of pragmatic labor can be accounted for by using Game Theory. In particular, we will look at the most simple games where information exchange is studied: signaling games.

This paper is organized as follows. We will introduce signaling games in section 2. In section 3, we will discuss how to account for Quantity_1 implicatures in signaling games. First, we will account for scalar implicatures by making use of the assumption that what speakers say must be *truthful* and *credible*. After that, we will discuss an analysis of Quantity_1 implicatures in terms of utility functions (and not just expected utility functions) that crucially depend on the probabilities involved. In section 4 we discuss various game theoretical approaches to account for Horn's division of pragmatic labor. According to the first, and more familiar approach, it is Pareto optimality that is crucial. The second, and new, approach makes use of Cho & Kreps' (1987) idea that messages can be 'equilibrium dominated' and of their 'Intuitive Criterion'.

2 Signaling games

Lewis (1969) defined the notion of a signaling game in order to explain the conventionalization of meaning of language without assuming any pre-existing

relation between messages and meanings.

A signaling game is a game of asymmetric information between a sender S and a receiver R. The sender observes the state S and R are in, while the receiver has to perform an action. The sender can try to influence the action taken by R by sending a message. This message, however, doesn't have an *a priori* given meaning. Let T be the set of (sender) states, F be the set of forms, or messages, and $\mathcal{A}$ be the set of actions such that $|T| = |\mathcal{A}| \leq |F|$. The sender will send a message/form in each state, a sender strategy σ is thus a function from T to F. The receiver will perform an action after hearing a message, a receiver strategy ρ is thus a function from F to $\mathcal{A}$. The utility functions U_{S} and U_{R} have real numbers as value, and take as arguments (i) the state t that S and R are in, (ii) the form $\sigma(t)$ sent by S in state t according to strategy σ, and (iii) the action $\rho(f)$ performed by R as a response to the received message f according to her strategy ρ. We assume that Nature picks the state according to some commonly known probability distribution P over T. The utility function for $i \in \{\mathrm{S}, \mathrm{R}\}$ is the expected utility relative to the probability distribution P over T:

$$\mathrm{EU}_i(\sigma, \rho) = \sum_{t \in T} P(t) \times U_i(t, \sigma(t), \rho(\sigma(t))).$$

A solution of the game is called a *Nash Equilibrium*. A Nash Equilibrium of a signaling game is a pair of strategies $\langle \sigma^*, \rho^* \rangle$ which has the property that neither the sender nor the receiver could increase his or her utility by unilateral deviation. Thus, $\langle \sigma^*, \rho^* \rangle$ is a Nash Equilibrium iff $\neg\exists\sigma : \mathrm{EU}_{\mathrm{S}}(\sigma, \rho^*) > \mathrm{EU}_{\mathrm{S}}(\sigma^*, \rho^*)$ and $\neg\exists\rho : \mathrm{EU}_{\mathrm{R}}(\sigma^*, \rho) > \mathrm{EU}_{\mathrm{R}}(\sigma^*, \rho^*)$.

The special thing about Lewisian signaling games is that all Nash equilibria survive many equilibrium refinements, including *Neologism proofness*, *perfect Bayesian equilibrium*, and the *Intuitive Criterion* which will be used in following sections of this paper.

As a small example, consider the *cheap talk* signaling game with only two states t_1, t_2, two messages $f_\varnothing, f$, and two actions a_1, a_2, where speaker and hearer have the same utility function U, and for all t_j and a_k, $U_i(t_j, f_\varnothing, a_k) = U_i(t_j, f, a_k) = 1$, if $j = k$, 0 otherwise. Obviously, the sender has four (pure) strategies. Because $f_\varnothing$ is the message of saying nothing, we can think of the receiver as having only 2 pure strategies (consistent with rational play). Furthermore, let $x = P(t_1) > P(t_2)$. Then, we have the following payoff matrix.

	t_1	t_2		$f_\varnothing$	f		ρ_1	ρ_2
σ_1	$f_\varnothing$	$f_\varnothing$	ρ_1	a_1	a_1	σ_1	x, x	x, x
σ_2	$f_\varnothing$	f	ρ_2	a_1	a_2	σ_2	x, x	$1, 1$
σ_3	f	$f_\varnothing$				σ_3	x, x	$0, 0$
σ_4	f	f				σ_4	x, x	$1-x, 1-x$

It is easy to see that this signaling game has four Nash equilibria: $\langle \sigma_1, \rho_1 \rangle$, $\langle \sigma_2, \rho_2 \rangle$, $\langle \sigma_3, \rho_1 \rangle$ and $\langle \sigma_4, \rho_1 \rangle$. As the reader can check, only in $\langle \sigma_2, \rho_2 \rangle$ does communication take place: meaning that in different states different messages are sent. Lewis calls such an equilibrium a 'signaling system', but we will call it a *separating equilibrium*. Lewis' (1969) main motivation for looking at signaling games is that in an equilibrium-play of the game, the messages can receive a meaning, although these meanings were not *a priori* assigned to these messages. The (descriptive) meaning of a message in an equilibrium like $\langle \sigma_2, \rho_2 \rangle$ is just the state, or set of states, in which the sender sends this message, e.g. the meaning of f in this equilibrium is $\sigma_2^{-1}(f) = \{t_2\}$. Notice that in the other equilibria of the game, the receiver simply ignores the messages being sent. Such equilibria are known as '*pooling* equilibria'. Lewis (1969) proposes that of all equilibria in a signaling game, only the separating ones can become a *convention*. A necessary condition for $\langle \sigma, \rho \rangle$ to be a separating equilibrium is that both σ and ρ are injective or one-to-one functions.

To change the example slightly (and preparing for section 4), we can assume that sending a message is *costly*. To implement this we can say that for all t_j and a_k, $U_i(t_j, f_\varnothing, a_k) = 1$, while $U_i(t_j, f, a_k) = 1 - \alpha$, if $j = k$, 0 otherwise (with $0 < \alpha < 1$). In this case, the game has only two Nash equilibria: the pooling equilibrium $\langle \sigma_1, \rho_1 \rangle$, and the separating equilibrium $\langle \sigma_2, \rho_2 \rangle$.

The above examples illustrates that not all Nash equilibria of a signaling game are separating. If we look at signaling games from an evolutionary point of view, however, it can be shown (Wärneryd, 1993) that only separating equilibria are evolutionarily stable. In this paper we won't go into the evolution of conventional meaning, and so won't go into evolutionary game theory.

3 Signaling games and Quantity$_1$ implicatures

3.1 Credibility and scalar implicatures

One of the reasons why we ended up with so many equilibria in Lewisian signaling games in the previous section where agents have identical utility functions was that the messages don't have an a priori given meaning. Farrell (1993) and others have shown that we can restrict this set of equilibria considerably, if we assume that messages have an exogenously given conventional meaning. Now

we can demand of messages that they should be sent *truthfully* and that equilibria should be *Neologism-Proof*. In this section I will show that in terms of these notions we can provide a game theoretic analysis of scalar implicatures.

There is an extensive literature in game theory that discusses under what circumstances a message is *credible*. Messages are seen as strategic moves to influence the hearer's actions. Intuitively, a statement is credible whenever the hearer has good reasons to believe what the speaker says. The statements discussed in game theory, *threats* and *promises* in particular, typically involve actions the speaker herself intends to carry out. For the purpose of this paper we don't need to go into this, and can limit ourselves to the discussion of credibility of messages in simple cheap talk signaling games introduced in the previous section. Even here we can limit ourselves to messages that simply state that the speaker is of a certain type. Let us say, for instance, that f_t is the message stating that the sender is of type t. In terms of an exogenously given semantic interpretation function $[\![\cdot]\!]$, it means that $[\![f_t]\!] = \{t\}$. Farrell (1988, 1992) and others (e.g. Stalnaker, 2006) would say that message f_t is credible in a certain signaling game if a sender of type t wants the receiver to believe that she is of type t. Based on some arguments of Aumann (1990), some have argued that in order for f_t to be credible, it also has to be the case that *only* a sender of type t wants the receiver to believe that the sender is of type t. If we abbreviate the optimal action for the receiver in case the sender is of type t by $BR_{\mathrm{R}}(t)$, we can formalize this notion of credibility as follows:

(1) Message f_t is **credible** for a sender if conditions (i) and (ii) hold:
(i) $U_{\mathrm{S}}(t, a) > U_{\mathrm{S}}(t, a')$ for $a \in BR_{\mathrm{R}}(t)$ and $a' \notin BR_{\mathrm{R}}(t)$.
(ii) $U_{\mathrm{S}}(t', a') > U_{\mathrm{S}}(t', a) : \forall t' \neq t \in T, a \in BR_{\mathrm{R}}(t)$ and $a' \in BR_{\mathrm{R}}(t')$.

The first part this definition says that if the sender is of type t, then she prefers f_t to be believed so that the receiver plays his optimal action a in t and not some other action a'. The second condition states that sender S wants her message to be believed (R plays a best response against it) only if she is of the type announced in the message. Thus, the message f_t is self-signaling when the sender wants it to be believed if and only if she is of type t. One can easily show that in a two-type situation, there exists a separating equilibrium iff there is at least one credible message f_t.

In order to study simple scalar implicatures, let us look at a two-type, two-message common interest cheap talk game where the messages do have a specific meaning. Let us assume that we have two types of situations: t_{all} and t_{sbna}, where t_{all} is a situation where John ate all of the cookies, while t_{sbna} is a situation where John ate some but not all of the cookies. The two messages are, of course, f_{all} with exogenously given semantic meaning ($[\![f_{all}]\!] = \{t_{all}\}$) that John ate all of the cookies, and f_{some} with semantic meaning ($[\![f_{some}]\!] =$

$\{t_{sbna}, t_{all}\}$) that John ate (at least) some of the cookies. Notice that f_{all} *semantically entails* f_{some}: $[\![f_{all}]\!] \subset [\![f_{some}]\!]$. Finally, we have the two actions a_{all} and a_{sbna}, such that $U(t_x, a_y) = 1$, if $x = y$ and 0 otherwise. Notice that in this setting, message f_{all} is credible:[1] the sender prefers a_{all} in t_{all} to a_{sbna}, and prefers a_{sbna} in the other situation. As mentioned above, in two-type situations we have a separating equilibrium iff there is at least one credible message. As in all two-type, two-message, two-action common interest cheap talk games, we have two separating equilibria, and, depending on the probability distribution also two or four pooling equilibria. To cut down this set of equilibria, we are going to make two assumptions: (i) in the equilibrium play of the game the messages must be used *truthfully*;[2] (ii) the equilibrium should be *Neologism Proof*. The first assumption can be thought of as an implementation of Grice's maxim of *Quality*. But the second one is crucial to account for scalar implicatures as well.

On our first requirement that in the equilibrium play of the game the messages that are used must be used *truthfully*, this means that half of the above mentioned equilibria disappear: the separating one where f_{all} in sent in t_{sbna}, and the pooling one(s) where the sender always sends f_{all}. This leaves us with two types of equilibria, if $P(t_{sbna}) > P(t_{all})$: (i) the pooling one(s) where f_{some} is always sent, and (ii) the separating one where f_{some} is sent in t_{sbna}, and f_{all} in t_{all}. Notice that we would be able to account for the scalar implicature in this example if we could explain why only the separating equilibrium where f_{all} is sent credibly is a sensible equilibrium. One way in which this can be done is by making use of evolutionary game theory.[3] But we don't really need to follow that path, if we make use of Farrell's (1993) requirement that equilibria be Neologism-Proof.

Farrell (1993) proposes that an equilibrium is *Neologism-Proof* if in no situation the speaker has an incentive to use an available (unused) credible message.[4] The intuition behind this notion is that if there exists a credible message f_t that is not used by the sender in the equilibrium play of the game,

[1]The question whether f_{some} is credible never arises according to our simple definition, because the semantic meaning of this message contains more than one type, $[\![f_{some}]\!] = \{t_{sbna}, t_{all}\}$.

[2]To implement this constraint, we require that $\forall t \in T$ and $\forall \sigma$, $t \in [\![\sigma(t)]\!]$. Though economists might find this an unnatural, or unmotivated, assumption, making such an assumption is uncontroversial among linguists and philosophers.

[3]Based on the earlier mentioned result of Wärneryd (1993) that only separating equilibria are evolutionarily stable.

[4]Cho & Kreps' (1987) *Intuitive Criterion* as used in section 4 of this paper is very close to Farrell's requirement. The only difference seems to be that the intuitive criterion relies on the cost of sending a certain message, not on the meaning. Thanks to the reviewer to make my earlier claim more precise.

and the sender of type t would be better off if she would have sent that message (and thus be believed, and acted upon by the receiver), then this equilibrium is not Neologism-Proof. For our simple common interest cheap talk game it is not difficult to see that no pooling equilibrium is Neologism-Proof, if $P(t_{sbna}) > P(t_{all})$. If there is a chance that the speaker is of type t_{all}, or is in situation t_{all}, it is always better for a sender of that type to send the credible message f_{all} instead of the weaker f_{some}. Because f_{all} can only be sent credibly by a speaker of type t_{all} this means that the equilibrium where t_{all} sends f_{all} and t_{sbna} sends f_{some} is the *unique* Neologism-Proof equilibrium of this game. But this means that although situation t_{all} is compatible with the *semantic* meaning of f_{some}, this message will only be sent in situation t_{sbna} in the unique Neologism-Proof equilibrium $\langle \sigma^*, \rho^* \rangle$ of the game: $\sigma^{*-1}(f_{some}) = \{t_{sbna}\}$. For this reason, we might call the inverse sender strategy $\sigma^{*^{-1}}$ of this unique equilibrium applied to a message f its *pragmatic* interpretation function. But then we see that 'John ate *some* of the cookies' pragmatically entails that John did not eat all of the cookies, i.e., the scalar implicature.

3.2 Quantity$_1$ implicatures in general

In this section, another analysis of standard scalar implicatures will be proposed. The analysis I will start with is motivated by Jäger's (manuscript) analysis of scalar implicatures, but is simpler and more general. It will be suggested afterwards that an appropriate generalization of this analysis can be used to account for more general Quantity$_1$ implicatures as well.[5]

To illustrate the new game theoretic treatment of Quantity implicatures, we will this time look at numerical expressions. Take a signaling game with 3 states, $T = \{t_1, t_2, t_3\}$, and three messages $F = \{$'one','two','three'$\}$. State t_i is the state where *exactly* i boys came to the party, while message 'n' has the semantic meaning that *at least* n boys came to the party. On this neo-Gricean 'at least' interpretation of numerals,[6] the meanings of the numeral expressions form an implication chain: $[\![$'three'$]\!] \subset [\![$'two'$]\!] \subset [\![$'one'$]\!]$. Just as

[5]To some readers this last sentence might conversationally implicate that the analysis of scalar implicatures proposed in the previous section cannot account for those more general Quantity$_1$ implicatures. This is indeed the case, but one can define a generalization of Farrell's condition of being Neologism-Proof — call it *Communication-Proof* – that an equilibrium has to satisfy. In terms of such a more general notion we could predict more general Quantity$_1$ implicatures, but the resulting analysis would be very similar to what will be proposed in this section.

[6]This assumption is controversial. Those who don't accept this assumption should think of other examples where the semantic meanings of the alternative expressions form a linear chain with respect to inference. The scales ⟨and, or⟩ and ⟨all, most, some⟩ would do if 'or' is read inclusively and the quantifiers 'all' and 'most' give rise to an existential presupposition.

in the previous section, we assume again that senders obey Grice's maxim of Quality and only say something that is true. Thus, if the speaker is in t_3 – the situation where exactly three boys came to the party – she could send all three messages, but if she is in a situation where only one boy came, t_1, she could only assert 'one'. This means that the sender can choose between six different strategies:

$$\begin{aligned}
\sigma_1 &= \{\langle t_1, \text{`one'}\rangle, \langle t_2, \text{`one'}\rangle, \langle t_3, \text{`one'}\rangle\} \\
\sigma_2 &= \{\langle t_1, \text{`one'}\rangle, \langle t_2, \text{`one'}\rangle, \langle t_3, \text{`two'}\rangle\} \\
\sigma_3 &= \{\langle t_1, \text{`one'}\rangle, \langle t_2, \text{`one'}\rangle, \langle t_3, \text{`three'}\rangle\} \\
\sigma_4 &= \{\langle t_1, \text{`one'}\rangle, \langle t_2, \text{`two'}\rangle, \langle t_3, \text{`one'}\rangle\} \\
\sigma_5 &= \{\langle t_1, \text{`one'}\rangle, \langle t_2, \text{`two'}\rangle, \langle t_3, \text{`two'}\rangle\} \\
\sigma_6 &= \{\langle t_1, \text{`one'}\rangle, \langle t_2, \text{`two'}\rangle, \langle t_3, \text{`three'}\rangle\}
\end{aligned}$$

The receiver's action is one of interpretation: he will assign an interpretation to each message. We assume that for each message f and receiver strategy ρ, $\rho(f) \subseteq [\![f]\!]$. If we also assume that $\rho(f) \neq \varnothing$ and that it also has to be *compact* (meaning that if $t_1, t_3 \in \rho(f)$, then it also has to be the case that $t_2 \in \rho(f)$), it means that the receiver can choose between 6 strategies:

$$\begin{aligned}
\rho_1 &= \{\langle \text{`one'}, \{t_1\}\rangle, \langle \text{`two'}, \{t_2\}\rangle, \langle \text{`three'}, \{t_3\}\rangle\} \\
\rho_2 &= \{\langle \text{`one'}, \{t_1\}\rangle, \langle \text{`two'}, \{t_2, t_3\}\rangle, \langle \text{`three'}, \{t_3\}\rangle\} \\
\rho_3 &= \{\langle \text{`one'}, \{t_1, t_2\}\rangle, \langle \text{`two'}, \{t_2\}\rangle, \langle \text{`three'}, \{t_3\}\rangle\} \\
\rho_4 &= \{\langle \text{`one'}, \{t_1, t_2\}\rangle, \langle \text{`two'}, \{t_2, t_3\}\rangle, \langle \text{`three'}, \{t_3\}\rangle\} \\
\rho_5 &= \{\langle \text{`one'}, \{t_1, t_2, t_3\}\rangle, \langle \text{`two'}, \{t_2\}\rangle, \langle \text{`three'}, \{t_3\}\rangle\} \\
\rho_6 = [\![\cdot]\!] &= \{\langle \text{`one'}, \{t_1, t_2, t_3\}\rangle, \langle \text{`two'}, \{t_2, t_3\}\rangle, \langle \text{`three'}, \{t_3\}\rangle\}
\end{aligned}$$

We assume that the message being sent is costless, and that the utility function is defined as follows: $U_{\mathrm{S}}(t, \rho(f)) = U_{\mathrm{R}}(t, \rho(f)) = P(t \,|\, \rho(f)) = 1/\,|\rho(f)|$ if $t \in \rho(f)$, 0 otherwise (where $|X|$ denotes the cardinality of set X). Now we can determine for each sender-receiver strategy combination its expected utility:

$\mathrm{EU}(\sigma, \rho)$	ρ_1	ρ_2	ρ_3	ρ_4	ρ_5	ρ_6
σ_1	$1/3$	$1/3$	$1/3$	$1/3$	$1/3$	$1/3$
σ_2	$1/3$	$1/2$	$1/3$	$1/2$	$2/9$	$7/18$
σ_3	$2/3$	$2/3$	$7/12$	$7/12$	$5/9$	$5/9$
σ_4	$2/3$	$4/9$	$1/2$	$1/3$	$5/9$	$5/18$
σ_5	$2/3$	$1/2$	$1/2$	$1/3$	$4/9$	$5/18$
σ_6	$\boxed{1}$	$5/6$	$5/6$	$4/6$	$7/9$	$11/18$

From this table we can easily see that the pair $\langle \sigma_6, \rho_1 \rangle$ is the only Nash equilibrium of this game. This is so because for all other combinations $\langle \sigma, \rho \rangle$, both players could always benefit if one of them would choose another strategy. Notice that our unique Nash equilibrium gives rise to a set of form-interpretation pairs according to which the number terms are given an 'exactly'-interpretation, which is the standard scalar implicature.

Unfortunately, this derivation is both not general enough and too specific. It is not general enough, because we don't want to limit ourselves to classical scalar implicatures based on *linear scales*. It is too specific, because we don't want to assume that the sender has *complete* information and knows in which state she is. Consider a case where the meanings of the alternative expressions are not linearly, but only partially ordered. Intuitively, when (2-b) is given as an answer to (2-a), the answer is interpreted as meaning that *only* John came to the party, while if (2-c) is given as answer, the interpretation is that either only John, or only Mary came:

(2) a. Who came to the party?
b. John came to the party.
c. John or Mary came to the party.

Suppose that John and Mary are the relevant persons for question (2-a) and that it is presupposed that somebody came. In that case, the sender's alternative answers consists naturally of the set {'John came', 'Mary came', 'John and Mary came', 'John or Mary came'}. This means that the meanings of these messages are not linearly, but only partially ordered. But to interpret those messages, we also have to take into account more situations than we have done above: not only should we look at the three (information-) states where (the speaker knows that) (i) only John came, (ii) only Mary came, and (iii) John and Mary came, but also ones where (the speaker knows that) (iv) only John or only Mary came, (v) at least John came, (vi) at least Mary came, (vii) at least John or Mary came. As suggested above, thinking of states here as states of the world is not good enough: we have to think of them as *information states*.

But once we think of states as information states, we can simply lift our strategies used in the previous example to more complicated ones, but leave the rest as it was. Thus, a speaker strategy is now a function from information states to messages, while a receiver strategy is a function from messages to *sets* of information states. The commonly known probability function P is now a function from information states to $[0, 1]$. We now define $U_{\mathrm{S}}(X, \sigma, \rho) = U_{\mathrm{R}}(X, \sigma, \rho)$ as $P(X|\rho(\sigma(X)))$, if $X \in \rho(\sigma(X))$, and $-n$ otherwise, with n sufficiently high (n is a real number, and $-n$ represents the penalty for violating Quality ($X \notin \rho(\sigma(X))$): the speaker should not say something that she doesn't know to be true). We won't go into a description of the

equilibria that we end up with now. But notice that in our above example we have 7 information states, or types, and only 4 messages. But this means that there are many equilibria that are *separating-like* in the sense that each message is sent in at least one information state. One can show (cf. de Jager & van Rooij, manuscript) that under certain probability distributions these separating-like equilibria are strict equilibria, and thus evolutionarily stable. At least one of those equilibria is of special importance for the analysis of Quantity implicatures. It is the one according to which a sender who sends message f knows that the semantic meaning of f is the case, but doesn't know that any stronger alternative message f' is true. The corresponding interpretation strategy is dubbed *Grice* in Van Rooij & Schulz (2004).

4 Horn's division of pragmatic labor

4.1 A game theoretic explanation of Horn's division

We have seen above how implicatures standardly based on Grice's first submaxim of Quantity can be given a game theoretic motivation. In this section we will show that these theories can also account for *Horn's division of pragmatic labor* – according to which an (un)marked expression (morphologically complex and less lexicalized) typically gets an (un)marked meaning –, which Horn (1984) claimed to follow from the interaction between both Gricean submaxims of Quantity, and the maxims of Relation and Manner. To illustrate, consider the following well-known example.

(3) a. John stopped the car.
b. John made the car stop.

We typically interpret the unmarked, or *light* message (3-a) as meaning stereotypical stopping, while the marked, or *costly* message (3-b) is interpreted as non-stereotypical stopping.

In the theory of costly signaling (Spence, 1973), costly messages can be used to turn games in which the preferences are not aligned to ones where they are. In this section we will see, however, that costly messages can also be used to indicate that the sender is of a remarkable type, which gives rise to a motivation of Horn's division of pragmatic labor. For our purposes, however, it is enough for the costs to be *nominal*, i.e., they never exceed the benefit of successful communication.[7]

[7] According to Blume et. al. (1993), this means that we are still in the realm of cheap talk signaling games.

Suppose we have 2 states, t_1 and t_2, and 2 messages, the *light* message f_l = (3-a), and the *costly* message f_c = (3-b). We assume that the semantic meanings of both expressions is the same, $[\![f_l]\!] = [\![f_c]\!] = \{t_1, t_2\}$. Let us assume, moreover, that according to the commonly known probability function P, $P(t_1) = \frac{3}{4} > \frac{1}{4} = P(t_2)$. The receiver has to choose between $\mathcal{A} = \{a_1, a_2\}$. The sender's utility function will be decomposable into a benefit and a cost function, $U_{\mathrm{S}}(t, f, a) = B(t, a) - C(f)$, while the receiver's utility function will just be equal to the benefit function, $U_{\mathrm{R}}(t, f, a) = B(t, a)$. As already indicated above, we assume that also for the sender it is always better to have successful communication with a costly message than unsuccessful communication with a cheap message. Thus, in contrast to the theory of costly signaling, we assume that the cost of sending a message can never exceed the benefit of communication. To assure this, we assume that $C(f_l) = 0$ and $C(f_c) = 0.2$, and adopt the following benefit function: $B(t_i, a_j) = 1$, if $i = j, 0$ otherwise. The sender- and receiver strategies are as before. The combination of sender and receiver strategies that gives rise to the bijective mapping $\{\langle t_1, f_l\rangle, \langle t_2, f_c\rangle\}$ is a Nash equilibrium of this game. And this (separating) equilibrium encodes Horn's division of pragmatic labor: the lighter message f_l expresses the stereotypical meaning t_1, while the non-stereotypical state t_2 is expressed by a heavier and more costly message f_c. Unfortunately, the game has two more equilibria: first there is the other separating equilibrium $\{\langle t_1, f_c\rangle, \langle t_2, f_l\rangle\}$ – where the lighter message denotes the non-stereotypical situation, and second there is also the pooling equilibrium where the sender always sends the lighter message, while the receiver maps all messages to a_1, which means that the message sent is ignored. We can conclude that on the present implementation the standard solution concept of game theory cannot single out the desired outcome, i.e., the first equilibrium.

Parikh (1992, 2001) argues that to account for this problem we should adopt another, and more fine-grained, solution concept. He observes that of the three equilibria mentioned above, the first one Pareto-dominates the others, and that for this reason the former should be preferred. But why should the Pareto-dominant equilibrium be selected? Van Rooij (2004) suggests that because Horn's division of pragmatic labor involves not only language use but also language organization, one should look at signaling games from an evolutionary point of view, and make use of those variants of evolutionary game theory that explain the emergence of Pareto-dominant solutions.

Although both proposals are appealing, I am not completely satisfied with either of them. Parikh's proposal to just select the Pareto-dominant Nash equilibrium seems somewhat ad hoc, while van Rooij's suggestion seems unnatural to explain those cases of Horn's division where it seems clear that it is only language *use* that counts. An obvious example is Grice's (1967) *Mrs*

T. produces a series of sounds closely corresponding the score of "Home Sweet Home". Because of the obvious alternative *Mrs T. sang "Home Sweet Home"*, the speaker wants to convey that there was something special with the singing. In the next section I will propose an alternative game theoretic explanation of Horn's division of pragmatic labor which is, I believe, more satisfying.[8]

4.2 Horn's division and the Intuitive Criterion

Suppose we start out with a decision problem of R. R knows that the actual state is either t_1 or t_2 and has to choose between a_1 and a_2, which have the following benefits in the states (where entry '$B(t, a) = a, b$' means that S's payoff is a, while R's payoff is b):

$B(t_i, a_j)$	a_1	a_2
t_1	1,1	0,0
t_2	0,0	1,1

Assume as before that $P_R(t_1) = \frac{3}{4} > \frac{1}{4} = P_R(t_2)$, it means that R is going to play a_1 because that is the action with the highest expected utility. Now suppose that there is another agent, S, who is known to have the same benefit function as R, but who knows in which state R and S are. Moreover, assume that it is known that S knows R's probability function P_R. S can send a message to influence R's decision, and thus we are formally involved in a signaling game. Suppose that S can send two messages: the 'empty' message $f_\varnothing$, which means doing nothing, and f_c which is a costly message. Both messages are compatible with t_1 and t_2. Assuming for concreteness that $C(f_\varnothing) = 0$, while $C(f_c) = 0.2$, this means that after f_c is sent, the utility functions are as follows:

send f_c

$U(t_i, f_c, a_j)$	a_1	a_2
t_1	0.8,1	-0.2,0
t_2	-0.2,0	0.8,1

This utility table is the same as the above benefit table, except that S's utilities are 0.2 utils lower because she sent costly message f_c. Notice that without doing anything, it is common knowledge that R would play a_1, because that has the highest expected utility. The question that arises now is what R is meant to infer from the observation that S chose to send the costly message f_c. 'Why did S, whom I know to be rational, play f_c given that this condemns her to a payoff of at most 0.8?, if I play my action with the highest expected utility'

[8]For a game theoretical treatment of something that is closely related to Horn's division of pragmatic labor, see Benz & van Rooij (to appear).

asks R upon observing f_c. 'Obviously, she is signaling to me that she doesn't want me to play a_1, presumable because she knows that she is in situation t_2 in order to get 0.8 utils, as opposed to the lower utility of 0!'. And since a_2 is R's best reply to t_2, this reasoning recommends R to respond to f_c by playing a_2.

This informal reasoning can be made formal by making use of Cho & Kreps' (1987) notion of a message being *equilibrium-dominated* for a sender of a particular type, and their *Intuitive Criterion*.

Let us think of the initial situation where only R faces a decision problem as a signaling game.[9] It is a very simple signaling game, because sender S has only one strategy: always saying nothing ($F = \{f_\varnothing\}$). The receiver can choose between a_1 and a_2. This 'signaling game' has only one solution: the sender always sends $f_\varnothing$, while the receiver plays the action with the highest expected utility: a_1. Now we are going to add another message: f_c. The semantic meaning of this message is still compatible with both t_1 and t_2, $[\![f_c]\!] = \{t_1, t_2\}$, but, intuitively, using such a message can still have an effect. To account for this intuition, we are going to make use of Cho & Kreps' (1987) definition of when a message is equilibrium-dominated. A message f is *equilibrium-dominated* for a sender of type t iff the sender's equilibrium payoff in t (denoted by $U^*_{\mathrm{S}}(t)$) is greater than the highest possible payoff she could receive if she sent f: $U^*_{\mathrm{S}}(t) > max_{a \in \mathcal{A}} U_{\mathrm{S}}(t, f, a)$. We will start out with the 'equilibrium' mentioned above, where the sender always sends $f_\varnothing$ and the receiver chooses a_1. Notice that the payoffs of senders in type t_1 and t_2 in this equilibrium are respectively $U^*_{\mathrm{S}}(t_1) = 1$ and $U^*_{\mathrm{S}}(t_2) = 0$, while $max_{a \in \mathcal{A}} U_{\mathrm{S}}(t_1, f_c, a) = 0.8$ and $max_{a \in \mathcal{A}} U_{\mathrm{S}}(t_2, f_c, a) = 0.8$. But this means that message f_c is equilibrium-dominated for a sender of type t_1, but not for one of type t_2. Now we propose that starting from the given equilibrium where nothing was said, we can rule out all sender strategies where the sender of type t_1 sends f_c. Thus, we require that in the new signaling game which also features message f_c, the sender would not use a strategy that assigns message f_c in situation t if f_c is equilibrium-dominated for a sender of type t. As a result, the sender can only choose between two strategies: (i) she always sends $f_\varnothing$, or (ii) sends $f_\varnothing$ in t_1 and f_c in t_2. Assuming that the receiver can still choose between all 4 pure strategies, we end up with two equilibria: (i) the *pooling* equilibrium where the speaker always says nothing (sends $f_\varnothing$ in both t_1 and t_2) while the receiver always plays a_1 (as a reaction to both $f_\varnothing$ and f_c), and (ii) the *separating* equilibrium where S sends $f_\varnothing$ in t_1 and f_c in t_2, while R uses strategy $\{\langle f_\varnothing, a_1\rangle, \langle f_c, a_2\rangle\}$.

[9] By starting with this pooling equilibrium, my proposal is closely related with Kris de Jaegher's (manuscript) evolutionary approach to Horn's division of Pragmatic labor. I profited from our discussion of my alternative way to proceed.

Cho & Kreps (1987) now argue that the pooling equilibrium is not natural: it doesn't satisfy their *Intuitive Criterion*. For Cho & Kreps' reasoning to go through, we have to make use of a somewhat stronger notion of equilibrium than the one due to Nash that we have used so far. This stronger notion is called a *Perfect Bayesian equilibrium* and is the standardly used solution concept in signaling games. It is stronger than a standard Nash equilibrium, because it requires the receiver to behave rational even in information states that will never be reached in the equilibrium play of the game, i.e. to behave *sequentially rational*. In particular, it requires of a pooling equilibrium with unused message f that the receiver would react to f with action a only if this action has the highest expected utility with respect to the belief state he would be in after receiving message f. Notice, now, that in order for our pooling equilibrium to be a perfect Bayesian equilibrium, it has to be the case that the conditional probability that R would assign to state t_1 if he received message f_c should be higher than $\frac{1}{2}$, because otherwise the speaker would not be sequentially rational. Formally there is no principle reason why this cannot be the case.[10] However, Cho & Kreps (1987) argue that having such conditional probabilities is unnatural. To implement this argument they state their *Intuitive Criterion* which says that the conditional probability that the receiver would assign to a type t after he received an out-of-equilibrium message f should be 0, if sending f is equilibrium-dominated for a sender of type t. In our example this means that according to this criterion, R's conditional probability of t_1 given f_c should be 0, which is inconsistent with the pooling equilibrium.

What this argument shows is that in order for a sender to send a message in the first place (even if it only has a nominal cost), it must be worthwhile to do so. One can argue that if one assumes that messages always have such (nominal) costs, this implements Grice's *maxim of Relevance*. This reasoning is also closely related to Horn's reasoning of why his division of pragmatic labor should hold, but by itself does not yet completely explain it. It does not yet completely explain it, because our reasoning above starts out from the pooling equilibrium where nothing was said. Indeed, I believe that not all cases of Horn's division *should* be explained by Cho & Kreps' Intuitive Criterion: the obvious distinction in meaning between *John went to jail* versus *John went to <u>the</u> jail* should, in our opinion, be accounted for in terms of language evolution, rather than language use. For other examples, however, I do believe the explanation is very intuitive. Consider again Grice's *Mrs T. produces a series of sounds closely corresponding the score of "Home Sweet*

[10]The pooling Nash equilibrium is also a pooling Bayesian equilibrium, because the S's strategy of this pooling equilibrium doesn't put any constraint on the conditional beliefs R could have if the unused message would have been sent.

Home", which is interpreted as saying that Mrs. T sang badly. I believe this sentence can be seen as saying *Mrs. T sang "Home Sweet Home" by producing a series of sounds closely corresponding to its score.* The crucial point of this re-wording is that this latter sentence can be seen as a conjunction of two messages, corresponding to a two-level interpretation on the receiver's side: first *Mrs. T. sang "Home Sweet Home"*, and then followed by *by producing a series of sounds closely corresponding to its score.*[11] Suppose that the actions correspond with the following interpretations: (1) normal singing, and (2) 'singing' badly. Because, by assumption, interpretation (1) is the most likely interpretation, this is what the receiver will do after receiving *Mrs. T. sang "Home Sweet Home"*. Now the receiver hears the second part of the sentence, and wonders what to make of this extra effort on the sender's side. Notice that if the sender knew that interpretation (1) was the correct one, i.e., if she was of type $t_{(1)}$, using extra effort would be equilibrium-dominated. Thus, the receiver will conclude via the Intuitive Criterion that the speaker must have meant something special with using the full sentence. In our context this means that he interprets the message as meaning that Mrs. T sang badly.

5 Conclusion

David Lewis (1969) used game theory to account for conventional meaning. The purpose of this paper was twofold: (i) to show that the theory of games could be used to account for conversational implicatures as well, but (ii) also to suggest that extra assumptions are required to do so. In this paper I used two types of refinements of a Nash equilibrium to account for some standard conversational implicatures: Neologism Proofness (in conjunction with the assumption that messages have a pre-existing meaning) to handle scalar implicatures, and the Intuitive Criterion to account for Horn's division of pragmatic labor. These refinements of the standard equilibrium notion are closely related, but by themselves they only rule out pooling equilibria. More is required to eliminate the undesired separating equilibria as well. For standard Quantity$_1$ implicatures, I used also Grice's maxim of Quality: the assumption that senders speak truthfully. For Horn's division of pragmatic labor, I made the extra assumption that messages should not be equilibrium dominated in a starting pooling equilibrium where only a cheap, or zero, message was used. The analyses of the two different types of implicatures are closely related, which suggests that an even more uniform treatment can be given. This, however, is something for the future.

[11] It is, of course, the two-level interpretation that is crucial for our analysis, not the particular re-wording that illustrates it.

References

[24] Atlas, J. and S. Levinson (1981), 'It-clefts, informativeness and logical form', In *Radical Pragmatics*, P. Cole (ed.), Academic Press, New York, 1-61.

[24] Aumann, R. (1990), 'Nash equilibria are not self-enforcing', In: J. Gabszewicz et al (eds.), *Economic Decision-Making: Games. Econometrics and Optimization*, Elsevier, Amsterdam, 201-206.

[24] Benz, A. and R. van Rooij (to appear), 'Optimal Assertions, and what they implicate. A uniform game theoretic approach', *Topoi*.

[24] Blume, A., Y.G. Kim and J. Sobel (1993), 'Evolutionary stability in games of communication', *Games and Economic Behavior*, **5**: 547-575.

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

[24] Cho, I.K. and D. Kreps (1987), 'Signaling games and stable equilibria', *Quarterly Journal of Economics*, **102**: 179-222.

[24] Farrell, J. (1988), 'Communication, coordination and Nash equilibrium', *Economic Letters*, **27**: 209-214.

[24] Farrell, J. (1993), 'Meaning and credibility in cheap-talk games', *Games and Economic Behavior*, **5**: 514-531.

[24] Gazdar, G. (1979), *Pragmatics*, London: Academic Press.

[10] Grice, H.P. (1967), 'Logic and conversation', *William James Lectures*, Harvard University, reprinted in *Studies in the Way of Words*, 1989, Harvard University Press, Cambridge, Massachusetts.

[24] Horn, L. (1972), *The semantics of logical operators in English*, PhD Thesis, Yale University.

[24] 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.

[24] Jager, T. de, and R. van Rooij (manuscript), 'Deriving Quantity Implicatures', Universiteit van Amsterdam.

[24] Jäger, G. (manuscript), 'Game dynamics connects semantics and pragmatics', University of Bielefeld.

[24] Jaegher, K. de (manuscript), 'The evolution of Horn's rule', Utrecht University.

[24] Levinson, S. (2000), *Presumptive Meanings. The Theory of Generalized Conversational Implicatures*, MIT Press: Cambridge, Massachusetts.

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

[24] Parikh, P. (1992), 'A game-theoretical account of implicature', in Y. Vardi (ed.), *Theoretical Aspects of Rationality and Knowledge: TARK IV*, Monterey, California.

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

[24] Rooij, R. van (2004), 'Signalling games select Horn strategies', *Linguistics and Philosophy*, **27**: 493-527.

[24] Rooij, R. van & K. Schulz (2004), 'Exhaustive interpretation of complex sentences', *Journal of Logic, Language and Information*, **13**: 491-519.

[24] Spence, A. M. (1973), 'Job market signaling', *Quarterly Journal of Economics*, **87**: 355- 374.

[24] Stalnaker, R. (2006), 'Saying and meaning, cheap talk and credibility', In: A. Benz, G. Jäger, R. van Rooij (eds.): *Game Theory and Pragmatics*, Palgrave, Macmillan, 83-100.

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