Evolution of conventional meaning and conversational principles

Robert van Rooy*

Abstract

In this paper we study language use and language organisation by making use of Lewisean signalling games. Standard game theoretical approaches are contrasted with evolutionary ones to analyze conventional meaning and conversational interpretation strategies. It is argued that analyzing successful communication in terms of standard game theory requires agents to be very rational and fully informed. The main goal of the paper is to show that in terms of evolutionary game theory we can motivate the emergence and self-sustaining force of (i) conventional meaning and (ii) some conversational interpretation strategies in terms of weaker and, perhaps, more plausible assumptions.

1 Introduction

We all feel that information transfer is crucial for communication. But it cannot be enough: although smoke indicates that there is fire, we wouldn't say that communication is taking place. Also not all transfer of information between humans counts as communication. *Incidental* information transfer should be ruled out. Intuitively, in order for an event to *mean* something else, *intentionality* is crucial. And indeed, Grice (1957) characterizes 'meaning' in terms of communicator's intentions. To mean something by x, speaker S must intend

(1) S's action x to produce a certain response a in a certain audience/receiver R;
(2) R to recognize S's intention (1);
(3) R's recognition of S's intention (1) to function as at least part of R's reason for R's response a.

The first condition says basically that we communicate something in order to influence the receiver's beliefs and/or behavior. However, for an act to be a communicative act, the response should be mediated by the audience's *recognition* of the sender's *intention*, i.e. condition 2. But also the recognition of the speaker's intention is not sufficient. To see what is missing, consider the following contrasting pair (Grice, 1957):

(1) (a) A policeman stops a car by standing in its way.
    (b) A policeman stops a car by waving.

*The research for this paper has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences (KNAW). I am grateful to Wiebe van der Hoek for inviting me to submit a paper to the current volume of KRA. I would like to thank Johan van Benthem, Matina Donaldson, Gerhard Jäger, Martin Stokhof, and two anonymous reviewers of this journal for critical comments on and useful suggestions to an earlier version of this paper.

Although in both examples the first two Gricean conditions are satisfied, we would say that only in (1b) some real communication is going on. The crucial difference between (1a) and (1b), according to Grice (1957), is that only in (1b) the audience's recognition of the policeman's intention to stop the car is *effective* in producing that response. In contrast to the case where he stands himself in the car's way, the policeman does not regard it as a foregone conclusion that his waving will have the intended effect that the driver stops the car, whether or not the policeman's intention is recognized.

To be able to characterize the contrast between (1a) and (1b) is important to characterize linguistic, or conventional meaning. The difference between (2a) and (2b) seems to be of exactly the same kind.

(2) (a) Feeling faint, a child lets its mother see how pale it is
(hoping that she may draw her own conclusion and help.)
(b) A child says to its mother, "I feel faint."

In contrast to (1a) and (2a), in the cases (1b) and (2b) an agent communicates something by means of a sign with a *conventional* meaning (cf. Lewis, 1969, pp. 152-159; Searle, 1969). But Grice did not really intend to characterize situations where agents intend to influence one another by making use of signals with a conventional meaning. He aimed to account for successful communication even without conventional ways of doing so.

According to Grice, *third*-order intentionality is required for communicative acts: the speaker *intends* the hearer to *recognize* that the speaker *wants* the hearer to produce a particular response. Strawson (1964) and Schiffer (1972) showed by means of some examples that this third-order intentionality is not enough. We can still construct examples where an agent wants her audience to recognize her intention in order to produce a certain effect, without it intuitively being the case that the speaker's action *means* the intended response: it can be that the speaker doesn't want her intention, that R performs the desired action, to become *mutually known*.[1] For an action to be called communicative, the action has to make the speaker's intention *common knowledge*.

In this paper we will study both conventional and non-conventional meaning in terms of signalling games as invented by David Lewis and developed further in economics and biology. However, we are going to suggest that in order to successfully communicate information we don't need as much rationality, higher-order intentionality or common knowledge as (explicitly or implicitly) required by Grice, Lewis, Schiffer, and others. Building on work of economists and biologists, we will suggest that evolutionary game theory can be used to account for the emergence and self-perpetual force of both *arbitrary semantic rules* and of general *functional pragmatic interpretation strategies*.

This paper is organized as follows. In section 2 the analysis of signalling in standard, or rational, game theory is examined. The standard problem here is that of *equilibrium selection* and in section 3 Lewis's (1969) *conventional* way of solving it is discussed, together with his motivation for why these conventional solutions are self-enforcing. In section 4

[1] Here is Schiffer's (1972) example. Suppose S wants R to think that the house he is thinking of buying is rat-infested. S decides to bring about this belief in R by letting loose a rat in the house. He knows that R is watching him and knows that R believes that S is unaware that R is watching him. S intends R to infer, wrongly, from the fact that he let the rat loose that he did so with the intention that R should see the rat, take the rat as 'natural' evidence, and infer therefrom that the house is rat-infested. S further intends R to realize that the presence of the rat cannot be taken as genuine evidence; but S knows that R will think that S would not be so contrive to get R to believe the house is rat-infested unless S had good reasons for thinking it was, and so intends R to infer that the house is rat-infested from the fact that S is letting the rat loose with the intention of getting R to believe that the house is rat-infested. In this example, S's action does intuitively not 'mean' that the house is rat-infested, although the Gricean conditions are all met. See Parikh (1991, 2001) for an interesting game-theoretical discussion of this example.

evolutionary game theory is used to provide an alternative motivation for why linguistic conventions remain stable and why some candidate conventions are more natural to emerge than others. According to this alternative motivation we don't need to assume as strong notions of rationality and (common) knowledge as Lewis does. In section 5 and 6 we argue that evolutionary signalling games can also be used to motivate why natural languages are organized and used in such an *efficient* but still *reliable* way. Reliability (Grice's maxim of quality) is tackled in section 5, efficiency (Grice's (1967) quantity and manner) in section 6. The paper ends with some conclusions and suggestions for further research.

2 Communication problems as signalling games

2.1 Signalling games

For the study of information exchange we will consider situations where a speaker has some relevant information that the hearer lacks. The simplest games in which we see this asymmetry are *signalling games*. A signalling game is a two-player game with a *sender*, s, and a *receiver*, r. This is a game of *private* information: The sender starts off knowing something that the receiver does not know. The sender knows the state $t \in T$ she is in but has no substantive payoff-relevant actions.[2] 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 message m drawn from some set M. The messages don't have a pre-existing meaning. The other player receives this signal, and then takes an action a drawn from a set $\mathcal{A}$. 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. For simplicity, we take T, M and $\mathcal{A}$ all to be finite. A pure *sender strategy*, S, is a (deterministic) *function* from states to signals (messages): $S \in [T \to M]$, and a pure *receiver strategy*, R, a (deterministic) function from signals to actions: $R \in [M \to \mathcal{A}]$. Mixed strategies (probabilistic functions, which allow us to account for ambiguity) will play only a minor role in this paper and can for the most part be ignored.

As an example, consider the following signalling game with two equally likely states: t and t'; two signals that the sender can use: m and m'; and two actions that the receiver can perform: a and a'. Sender and receiver each have now four (pure) strategies:

Sender :

	t	t'
S_1	m	m'
S_2	m	m
S_3	m'	m
S_4	m'	m'

Receiver :

	m	m'
R_1	a	a'
R_2	a'	a
R_3	a	a
R_4	a'	a'

To complete the description of the game, we have to give the *payoffs*. The payoffs of the sender and the receiver are given by functions U_s and U_r, respectively, which (for the moment) are elements of $[T \times \mathcal{A} \to \mathbf{R}]$, where $\mathbf{R}$ is the set of reals. Just like Lewis (1969) we assume (for the moment) that sending messages is costless, which means that we are talking about *cheap talk* games here.

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

Coming back to our example, we can assume, for instance, that the utilities of sender and receiver are in perfect alignment – i.e., for each agent i, $U_i(t,a) = 1 > 0 = U_i(t,a')$ and $U_i(t',a') = 1 > 0 = U_i(t',a)$.[3]

An *equilibrium* of 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, $U_s^*(t,S,R)$, in any state t:

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

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, m, 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 in which the speaker could have sent message m. Let us define S_t to be the *information state* (or information set) the receiver is in after the sender, using strategy S, sends her signal in state t, i.e. $S_t = \{t' \in T : S(t') = S(t)\}$.[4] With respect to this set, we can determine the (expected) utility of receiver strategy R in information state S_t, which is R's expected utility in state t when the sender uses strategy S, $U_r^*(t,S,R)$ (where $P(t'|S_t)$ is the conditional probability of t' given S_t):

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

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:[5]

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

As can be checked easily, our game has 6 Nash equilibria: $\{\langle S_1,R_1\rangle, \langle S_3,R_2\rangle, \langle S_2,R_3\rangle, \langle S_2,R_4\rangle, \langle S_4,R_3\rangle, \langle S_4,R_4\rangle\}$. This set of equilibria depends on the receiver's probability function. If, for instance, $P(t) > P(t')$, then $\langle S_2,R_4\rangle$ and $\langle S_4,R_4\rangle$ are no equilibria anymore: it is always better for the receiver to perform a.

In signalling games it is assumed that the messages have no pre-existing meaning. However, it is possible that meanings can be associated with them due to the sending and receiving strategies chosen in equilibrium. If in equilibrium the sender sends different messages in different states and also the receiver acts differently on different messages, we can say with Lewis (1969, p. 147) that the equilibrium pair $\langle S,R\rangle$ fixes meaning of expressions in the following way: for each state t, the message $S(t)$ means either $S_t = \{t' \in T | S(t') = S(t)\}$ (in the case that the sentence is used *indicatively*) or $R(S(t))$ (if the sentence is used *imperatively*).[6]

[3]This assumption allows Hurford (1989), Oliphant (1996), Nowak & Krakauer (1999) and others to represent sender and receiver strategies by convenient transmission and reception matrices.

[4]Throughout the paper we will assume that communication is 'noiseless'. Although interesting possibilities arise when we give it up – as shown by Nowak & Krakauer (1999), it typically leads to more distinctive, or discrete, signals –, we will for simplicity assume that the receiver has no difficulties to perceptually distinguish the message being sent. Detecting its meaning is already hard enough.

[5]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.

[6]In our Lewisean games the two moods always come together. Searle (1969, pp. 42-50) argued that the concept of 'meaning' can only be applied to illocutionary effects, not to perlocutionary ones. In the main text I will limit myself normally only to 'indicative' meaning, which might well be in accordance with Searle's proposal.

Following standard terminology in economics (e.g. Crawford & Sobel, 1982), let us call $\langle S, R\rangle$ a (fully) *separating* equilibrium if there is a one-to-one correspondence between states (meanings) and messages, i.e., if there exists a bijection between T and M. Notice that among the equilibria in our example, two of them are separating: $\langle S_1, R_1\rangle$ and $\langle S_3, R_2\rangle$.

2.2 Requirements for successful communication

In the introduction we have seen that according to Schiffer an action only counts as being communicative if it makes the speaker's intention *common knowledge*. It can be argued that this common-knowledge requirement is met if a game has a *unique solution*. It is well-known (e.g. Osborne & Rubinstein, 1994) that in order for a strategy pair to be a Nash equilibrium, both the strategies that the agents can play and the preferences involved have to be common knowledge. Moreover, it is required that it is common knowledge that both agents are rational selfish payoff optimizers. If then, in addition, a particular signalling game has only one (Nash) solution, it seems only reasonable to claim that in that case the speaker's intention becomes common knowledge after she sent a particular signal.[7]

Thus we might claim communication to take place by sending message m in such a game if and only if (i) the game has a (partly or fully) *separating* equilibrium in which message m is sent; and (ii) this is the *unique* solution of the game.[8] The first condition is prominent in the economic and biological literature on signalling games. The second, uniqueness, condition plays an important role in Schelling (1960), Lewis (1969), and Clark (1996) to solve coordination problems and is stressed in the work of Parikh (1991, 2001) on situated communication. The following example shows that in case of non-arbitrary signals this uniqueness condition is sometimes indeed unproblematically satisfied.

Consider the following abstract situation. There are two kinds of situations: t, the default case where there is no danger; and t' where there is danger. The sender knows which situation is the case, the receiver does not. We might assume for concreteness that it is commonly known between sender and receiver that $P(t) = 0.8 > 0.2 = P(t')$. In the normal situation, t, the sender doesn't send a message, but in the other case she might. The message will be denoted by m, while not sending a message will be modelled as sending ϵ. The receiver can perform two kinds of actions: the default action a (which is like doing nothing); and action a'. This latter action demands effort from the receiver, but is the only appropriate action in the case that there is danger. It doesn't harm the sender if it is done if there is no danger (the sender is ambivalent about the receiver's response in t). One way to describe this situation is by assuming the following (also commonly known) utility functions:

$$\begin{array}{llll} U_s(t,a) = 5, & U_s(t,a') = 5, & U_s(t',a) = -50, & U_s(t',a') = 50 \\ U_r(t,a) = 6, & U_r(t,a') = 0, & U_r(t',a) = -10, & U_r(t',a') = 10 \end{array}$$

The strategies are as expected: S is just a function from t' to $\{m, \epsilon\}$, where ϵ is the empty message that is always sent in t; while R is a function from $\{\epsilon, m\}$ to $\{a, a'\}$. Thus, we have the following strategies

[7] See also Lewis's (1969, pp. 152-159) proof that if a signal is used conventionally (in Lewis's sense), the Gricean (1957) requirements for non-natural meaning are met (though not necessarily the other way around).

[8] The second condition is a little bit too strong. It is enough to require that all solutions of the game assign to m the same meaning.

Sender :

	t	t'
S_1	ϵ	ϵ
S_2	ϵ	m

Receiver :

	ϵ	m
R_1	a	a
R_2	a'	a'
R_3	a	a'
R_4	a'	a

On assuming that $P(t) = 0.8$, we receive the following payoff tables (for $U_i^*(\cdot, S, R)$):

t :

	R_1	R_2	R_3	R_4
S_1	**5,2.8**	5,2	**5,2.8**	5,2
S_2	**5,6**	5,2	**5,6**	5,2

t' :

	R_1	R_2	R_3	R_4
S_1	-50,2.8	50,2	-50,2.8	**50,2**
S_2	-50,-10	**50,10**	**50,10**	-50,-10

These payoff tables show that our game has exactly one Nash equilibrium: $\langle S_2, R_3 \rangle$, because only this strategy pair is an equilibrium (is boxed) in both states. Because in this game the unique-solution requirement is satisfied, we can be sure that communication is successful: If the sender sends m, the receiver will figure out that he is in situation t' and should perform a'.

Our game has exactly one Nash equilibrium in which meaningful communication is taking place because the sender has an incentive to influence the hearer and the receiver has no dominating action. If either the sender sees no value in sending information, or the receiver counts any incoming information as valueless for his decision, a signalling game will (also) have so-called 'pooling' equilibria, in which the speaker always sends the same message, and 'babbling' equilibria where the receiver ignores the message sent by the speaker and always 'reacts' by choosing the same action. In such equilibria no information exchange is taking place. One reason for why a receiver ignores the message sent might be that he cannot (always) take the incoming information to be *credible*.

A message is not credible if an individual might have an incentive to send this message in order to deceive her audience. In an important article, Crawford & Sobel (1982) show that the amount of credible information exchange in (cheap talk) games depends on how far the preferences of the participants are aligned.[9] However, this doesn't mean that in all those cases successful communication takes place when the sender sends a message. The unique solution requirement has to be satisfied as well, for otherwise sender and receiver are still unclear about the strategy chosen by the other conversational participant. Above we saw that in some cases such a unique solution is indeed possible. The example discussed in section 2.1 suggests, however, that in signalling games in which messages have no pre-existing meaning, the satisfaction of the uniqueness condition is the exception rather than the rule.[10] Even limiting ourselves to *separating* equilibria won't do. The problem is that that game has *two* such equilibria: $\langle S_1, R_1 \rangle$ and $\langle S_3, R_2 \rangle$.[11] How is communication possible in such a situation?

[9] See, among many others, van Rooy (2003) for more discussion.

[10] Parikh (1991, 2001) assumes a stronger solution concept (Pareto optimality) than that of a Nash equilibrium as I assume here (for a definition of Pareto optimality, see Tuyls et al (this volume)) . With the help of this concept, more games satisfy the unique-solution-condition (though not the one discussed in 2.1). In van Rooy (in press) I argue against the use of this concept in rational game theory, but show that the emergence of Pareto optimal solutions can be explained if games are thought of from an evolutionary point of view. See also section 6.1 of this paper for more discussion.

[11] Lewis (1969, p. 133) calculates that similar signalling problems with m states and n signals have $\frac{n!}{(n-m)!}$ separating equilibria.

3 A language as a conventional signalling system

3.1 Conventions as rationally justified equilibria

Above we assumed that the agents had no real prior expectations about what the others might do. Consider a simple symmetric two-person coordination game where both have to choose between a and b; if they both choose the same action they earn 1 euro each and nothing otherwise. If both take either of the other's actions to be equally likely (i.e., there are no prior expectations yet), the game has two (strict) Nash equilibria: $\langle a, a\rangle$ and $\langle b, b\rangle$. Things are different if each player takes it to be more likely that the other player will choose, say, a. In that case, both have an incentive to play a themselves as well: the expected utility of playing a is higher than that of playing b. But it is not yet a foregone conclusion that both also actually *should* play a: the first agent might believe, for instance, that the other player doesn't believe that the first will play a and she doesn't take the second player to be rational. That is, the beliefs of the agents need not be coherent (with themselves, or/and with each other). In that case, the first agent might have an incentive not to play a. This won't happen, of course, when the beliefs of the two agents and their rationality are *common knowledge* (or common belief). In that case, action combination $\langle a, a\rangle$ is the only Nash equilibrium of the game.

In the light of the above discussion, Lewis (1969) gave a straightforward answer of how agents coordinate on a particular signalling equilibrium: it is based on the commonly known expectation that the other will do so and each other's rationality. Confronted with the recurrent coordination problem of how to successfully communicate information, the agents involved take one of the equilibria to be the conventional way of solving the problem. This equilibrium $\langle S, R\rangle$ can be thought of as a *signalling convention*; a coding system that conventionally relates messages with meanings.

According to Lewis (1969), a signalling convention is a partially *arbitrary* way to solve a *recurrent* signalling situation of which it is commonly assumed by both agents that the other conforms to it. Moreover, it has to be commonly known that the belief that the other conforms to it, means that both have a good and decisive reason to conform to it themselves, and will want the other to conform to it as well. A *linguistic convention* is then defined as a generalization of such a signalling convention, where the problem is how to resolve a recurrent coordination problem to communicate information in a larger community.

We would like to explain a convention's (i) *emergence* and (ii) its *self-perpetuating* force.

Thinking of a convention as a special kind of equilibrium concept of rational game theory gives Lewis a straightforward explanation of why a convention is *self-sustaining*. Notice that the condition requiring that the belief that the other conforms to it means that both have a *good and decisive* reason to conform to it themselves is stronger than that of a Nash equilibrium: it demands that if the other player chooses her equilibrium strategy, it is *strictly* best (i.e., payoff-maximizing) for an agent to choose the equilibrium strategy too. Thus, according to Lewis, a convention has to be a *strict* Nash equilibrium.[12] Strict

[12] In Lewis (1969, pp. 8-24) an even stronger requirement is made. It is required for every player i that if all the other players choose their equilibrium strategies, it is best *for every player* that i chooses her equilibrium strategy too. This equilibrium concept is called a *coordinating equilibrium* and in terms of it Lewis wants to rule out the possibility that an equilibrium in games of (partly) conflicting interests (e.g. the game of Chicken) can be called a convention (and explain why conventions tend to become *norms* (ibid, pp. 97-100)). Vanderschraaf (1995) argued – convincingly we think – that there is a more natural way to rule out equilibria in such games to be called conventions: conventions have to satisfy a *public intentions criterion*, PIC: At a convention, each player will desire that her choice of strategy is common knowledge among all agents engaged in the game. Vanderschraaf also extends Lewis's notion of a convention by thinking of it as

equilibria in rational game theory are sustained simply by *self-interest*: if one expects the other to conform to the convention, unilateral deviation makes one (strictly) worse off.[13]

The notion of a strict equilibrium is stronger than the standard Nash equilibrium concept used in game theory. In terms of it we can explain why some equilibria are unlikely candidates for being conventions. Recall that the game discussed in section 2.1 had 6 Nash equilibria: $\{\langle S_1, R_1\rangle, \langle S_3, R_2\rangle, \langle S_2, R_3\rangle, \langle S_2, R_4\rangle, \langle S_4, R_3\rangle, \langle S_4, R_4\rangle\}$. We have seen that only the first two are separating: different messages are sent in different states such that there exists a 1-1 correspondence between meanings and messages. According to Lewis's (1969) definition of a convention, only these separating equilibria are appropriate candidates for being a convention, and he calls them *signalling systems*.

In the previous section we were confronted with what game theorists call the problem of *equilibrium selection*. Which of the (separating) equilibria of the game should the players coordinate on to communicate information? Lewis proposed to solve this problem by assuming that one of those equilibria is a convention. Which one of the (separating) equilibria should be chosen to communicate information is, in some sense, *arbitrary*, and it is this fact that makes both separating equilibria $\langle S_1, R_1\rangle$ and $\langle S_3, R_2\rangle$ equally appropriate candidates for being a convention (for solving the recurrent coordination problem at hand). In some sense, however, Lewis's solution just pulls the equilibrium selection problem back to another level: How are we to explain which of these regularities comes about? Two natural ways to establish a convention are explicit agreement and precedence. But for *linguistic* conventions the first possibility is obviously ruled out (at least for a first language), while the second possibility just begs the question. Following Lewis's (1969) proposal of how to solve coordination problems, this leaves *salience* as the last possibility. A salient equilibrium is one with a distinguishing psychological quality which makes it more compelling than other equilibria. With Skyrms (1996), we find this a doubtful solution for linguistic conventions: why should one of the separating equilibria be more salient than the other? But then, how can one signalling equilibrium be selected without making use of the psychological notion of salience?

Not only is Lewis's account of equilibrium selection problematic, his explanation of the self–perpetuating force of signalling equilibria isn't completely satisfactory either. His explanation crucially makes a strong *rationality* assumption concerning the agents engaged in communication. Moreover, as for all equilibria concepts in standard game theory, a lot of *common knowledge* is required; the rules of the game, the preferences involved, the strategies being taken (i.e., lexical and grammatical conventions), and the rationality of the players must all be common knowledge.[14] Though it is unproblematic to accept that the strong requirements for being common knowledge can be met for simple pieces of information, with Skyrms (1996) we find it optimistic to assume that they are met for complicated language games played by large populations.

3.2 Natural Conventions

Lewis (1969) admits that agents can conform to a signalling (or linguistic) convention without going through the explicit justification of why they should do so, i.e. without taking

a correlated equilibrium. In this way, also some 'unfair' equilibria (as in the Battle of Sexes game) are ruled out as candidates for conventions. We won't come back to Vanderschraaf's PIC or his latter extension in this paper.

[13]The strength of this self-sustaining force of an equilibrium depends crucially on the strength of the expectations on what others will do. With weaker expectations, 'safer' equilibria are more attractive.

[14]A proposition p is common knowledge for a set of agents if and only if (i) each agent i knows that p, and (ii) each agent j knows that each agent i knows that p, each agent k knows that each agent j knows that each agent i knows that p, and so on.

into account what the others are supposed to do, or what they expect the agent herself to do. Agents can use a signalling system simply out of *habit* and they might have learned this habit just by *imitating* others. These habits are self-perpetuating as well: if each individual conforms to the signalling convention out of habit, there is no reason to change one's own habit. Still, Lewis argues that rationality is important: the habit has a rational justification. That might be so, but, then, not any justification for a habit is necessarily the correct *explanation* of why the habit *is* followed. Although rationality considerations arguably play a crucial role in learning and in following the conventions of one's *second* language, this is not so clear when one learns and speaks one's mother's tongue. But if that is so, the higher-order intentions that Grice, Lewis, and others presuppose for successful communication are perhaps not as crucial as is standardly assumed.

For signal m to mean a, a receiver doesn't always have to do a because of its conscious 'recognition of the sender's intention for it to do a'. According to a *naturalistic* approach towards meaning (or intentionality) – as most forcefully defended by Millikan (1984) in philosophy and also adopted by biologists thinking of animal communication as Maynard Smith & Harper (1995) – all that is needed for a signal to 'mean' something is that the sender-receiver combination $\langle S, R\rangle$ from which this message-meaning pair follows must be *selected* for by the force of evolution. In this way – as stressed by Millikan (1984) – a potential distinction is made not between human and animal communication, but rather between animal (including human) communication and 'natural' relations of indication. In distinction with the dances of honeybees to indicate where there is nectar to be found, smoke is not selected for by how well it indicates fire.[15] Just as Crawford & Sobel (1982) show that (cheap talk) communication is possible only when signalling is advantageous for both the sender and the receiver, in the same way it is guaranteed that for a signalling pair to be *stable*, there must be a *selective advantage* both (i) in attending and responding to the signals and (ii) in making them. This seems to be a natural reason for why a signalling convention has *normative* features as well. Evolutionary game theory (EGT) is used to study the notion of stability under selective pressures. Where traditional game theory is a normative theory with hyperrational players, EGT is more descriptive. It starts from a realistic view of the world, where players are neither hyperrational, i.e., are limited in their computational resources in their ability to reason, nor fully informed.

4 Stability and evolution in game theory

Lewis (1969) proposed to explain why linguistic conventions are self-sustaining in terms of rational game theory. To do so, he was forced to make very strong assumptions concerning agents' rationality and (common) knowledge. This suggests that we should look for another theoretical underpinning of the self-sustaining force of signalling conventions. Above, we have seen that perhaps an (unconscious) mechanism like habit is an at least as natural reason for a linguistic convention to remain what it is. In this section we will show that by adopting an evolutionary stance towards language, such a simpler mechanism might be enough for linguistic conventions to be stable. Our problem, i.e. which signalling conventions are self-sustaining, now turns into a problem of which ones are *evolutionarily stable*, i.e., *resistant* to variation/mutation.

In section 3.1 we have thought of a sender-receiver strategy pair $\langle S, R\rangle$ as a signalling convention to resolve a recurrent coordination problem to communicate information. We

[15] In the information theoretic account of content as developed by Dretske (1981) and others, our concept of evolution is replaced by that of *learning*. Though the two are not the same, they are related (cf. the paper of Tuyls et al. in this volume): both take the *history* of the information carrying device to be crucial.

assumed that all that matters for all players was successful communication and that the preferences of the agents are completely aligned. A simple way to assure this is to assume that $A = T$ and that all players have the following utility function:

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

Implicitly, we still assumed that individuals have fixed roles in coordination situations: they are always either a sender or a receiver. In this sense it is an *asymmetric* game. It is natural, however, to give up this assumption and turn it into a *symmetric* game: we postulate that individuals can take both the sender- and the receiver-role. Now we might think of a pair like $\langle S, R\rangle$ as a *language*. We abbreviate the pair $\langle S_i, R_i\rangle$ by L_i and take $U_s(t, L_i, L_j) = U(t, R_j(S_i(t)))$ and $U_r(t, L_i, L_j) = U(t, R_i(S_j(t)))$.

Consider now the symmetric strategic game in which each player can choose between finitely many languages. On the assumption that individuals take both the sender and the receiver role half of the time, the following utility function, $\mathcal{U}(L_i, L_j)$, is natural for an agent with strategy L_i who plays against an agent using L_j (where $EU_i(L, L')$ denotes the expected utility for i to play language L if the other participant plays L', i.e. $\sum_t P(t) \times U_i(t, L, L')$).

$$\begin{array}{lll} \mathcal{U}(L_i, L_j) & = & [\frac{1}{2} \times (\sum_t P(t) \times U_s(t, L_i, L_j))] + [\frac{1}{2} \times (\sum_t P(t) \times U_r(t, L_i, L_j))] \\ & = & \frac{1}{2} \times (EU_s(L_i, L_j) + EU_r(L_i, L_j)) \end{array}$$

Now we say that L_i is a (Nash) equilibrium of the language game iff $\mathcal{U}(L_i, L_i) \geq \mathcal{U}(L_i, L_j)$ for all languages L_j. It is straightforward to show that language L_i is a (strict) equilibrium of the (symmetric) language game if and only if the strategy pair $\langle S_i, R_i\rangle$ is a (strict) equilibrium of the (asymmetric) signalling game.

Under what circumstances is language L evolutionarily stable? Thinking of strategies immediately as languages, standard evolutionary game theory (see Maynard Smith, 1982; Weibull 1995, and others) gives the following answer.[16] Suppose that all individuals of a population use language L, except for a fraction ϵ of 'mutants' which have chosen language L'. Assuming random pairing of strategies, the expected utility, or *fitness*, of language $L_i \in \{L, L'\}$, $\mathcal{EU}^\epsilon(L_i)$, is now:

$$\mathcal{EU}^\epsilon(L_i) = (1 - \epsilon)\mathcal{U}(L_i, L) + \epsilon\mathcal{U}(L_i, L')$$

In order for mutation L' to be driven out of the population, the expected utility of the mutant need to be less than the expected utility of L, i.e., $\mathcal{EU}^\epsilon(L) > \mathcal{EU}^\epsilon(L')$. To capture the idea that mutation is extremely rare, we require that a language is *evolutionarily stable* if and only if there is a (small) number n such that $\mathcal{EU}^\epsilon(L) > \mathcal{EU}^\epsilon(L')$ whenever $\epsilon < n$. Intuitively, the larger n is, the 'more stable' is language L, since larger 'mutations' are resisted.[17] As is well-known (e.g. Maynard Smith, 1982), this definition comes down to Maynard Smith & Price's (1973) concept of an evolutionarily stable strategy (ESS) for our language game.

[16] Although evolutionary game theory was first used to model replication through genetic inheritance, it can and has been successfully applied to the evolution of social institutions as well, where replication goes by imitation, memory and education. For linguistic conventions we think of evolution in cultural rather than genetic terms. Fortunately, as shown by Tuyls et al. (this volume) and others, there are at least some learning mechanisms (e.g. multi-agent reinforcement learning, social learning) that provide a justification for our use of the replicator dynamics that underlies the evolutionarily stability concept we use, in the sense that (in the limit) they give rise to the same dynamic behavior. Also the Iterated Learning Mechanism used by Hurford, Kirby and associates shows at least in some formulations a great similarity with that of evolutionary games.

[17] The fact that linguistic conventions need not be resistant to larger mutations enables the theory to allow for language change from one 'stable' state to another.

Definition 1 *(Evolutionarily Stable Strategy, ESS)*
Language L is Evolutionarily Stable in the language game with respect to mutations if

1. *$\langle L, L\rangle$ is a Nash equilibrium, and*
2. *$\mathcal{U}(L', L') < \mathcal{U}(L, L')$ for every best response L' to L for which $L' \neq L$.*

We see that $\langle L, L\rangle$ can be a Nash equilibrium without L being evolutionarily stable (see Tuyls et al. (this volume) for more discussion). This means that the standard equilibrium concept in evolutionary game theory is a *refinement* of its counterpart in standard game theory (see Tuyls et al. (this volume) for more on the relation between the different equilibrium concepts). As it turns out, this refinement gives us an alternative way from Lewis (1969) to characterize the Nash equilibria that are good candidates for being a convention.

In an interesting article, Wärneryd (1993) proves the following result: For any sender-receiver game of the kind introduced above, with the same number of signals as states and actions, a language $\langle S, R\rangle$ is evolutionarily stable if and only if it is a (fully) separating Nash equilibrium.[18] In fact, this result follows immediately from more general game theoretical considerations. First, it follows already directly from the definition above that being a *strict* Nash equilibrium is a *sufficient* condition for being an ESS. Given that in our asymmetric cooperative signalling games the separating equilibria are the strict ones, a general result due to Selten (1980) – which states that in asymmetric games all and only the strict equilibria are ESS – shows that this is also a *necessary* condition. Thus we have the following

Fact 1 *(Wärneryd (and Selten)) In a pure coordination language game, L is an ESS if and only if $\langle L, L\rangle$ is a separating Nash equilibrium.*

In this way Wärneryd (and Selten) has given an appealing explanation of why Lewisean signalling systems are self-sustaining without making use of a strong assumption of rationality or (common) knowledge. But this is not enough for the evolutionary stance to be a real alternative to Lewis's approach towards conventions. It should also be able to solve the *equilibrium selection* problem. Which of the potential candidates is actually selected as the convention? As it turns out, also this problem has an appealing evolutionary solution, if we also take into account the *dynamic process* by which such stable states can be reached.

Taylor & Jonker (1978) defined their *replicator dynamics* to provide a continuous dynamics for evolutionary game theory. It tells us how the distribution of strategies playing against each other changes over time.[19] A *dynamic equilibrium* is a fixed point of the dynamics under consideration. A dynamic equilibrium is said to be *asymptotically stable* if (intuitively) a solution path where a small fraction of the population starts playing a mutant strategy still converges to the stable point (for more discussion, see Tuyls et al. (this volume) and

[18] This result doesn't hold anymore when there are more signals than states (and actions). We will have some combinations $\langle S, R_i\rangle$ and $\langle S, R_j\rangle$ which in equilibrium give rise to the same behavior, and thus payoff, although there will be an unused message m where $R_i(m) \neq R_j(m)$. Now these combinations are separating though not ESS. Wärneryd defines a more general (and weaker) evolutionary stability concept, that of an evolutionarily stable *set*, and shows that a strategy combination is separating if and only if it is an element of such a set.

[19] For our language game this can be done as follows: On the assumption of random pairing, the *expected utility*, or fitness, of language L_i at time t, $\mathcal{EU}_t(L_i)$, is defined as:
$\mathcal{EU}_t(L_i) = \sum_j P_t(L_j) \times \mathcal{U}(L_i, L_j)$
The expected, or average, utility of a population of languages $\mathbf{L}$ with probability distribution P_t is then:
$\mathcal{EU}_t(\mathbf{L}) = \sum_{L \in \mathbf{L}} P_t(L) \times \mathcal{EU}_t(L)$
The *replicator dynamics* (for our language game) is then defined as follows:
$\frac{\mathrm{d}P(L)}{\mathrm{d}t} = P(L) \times (\mathcal{EU}(L) - \mathcal{EU}(\mathbf{L}))$.

references therein). Asymptotic stability is a refinement of the Nash equilibrium concept. And one that is closely related with the concept of ESS. Taylor & Jonker (1978) show that every ESS is asymptotically stable. Although in general it isn't the case that all asymptotically stable strategies are ESS, on our assumption that a language game is a *cooperative* game (and thus *doubly* symmetric)[20] this is the case. Thus, we have the following

Fact 2 *A language L is an ESS in our language game if and only if it is asymptotically stable in the replicator dynamics.*

The 'proof' of this fact follows immediately from some important more general results provided by Weibull (1995, section 3.6). First, he shows that Fisher's (1930) so-called *fundamental theorem of natural selection* – according to which evolutionary selection induces a monotonic increase over time in the *average* population fitness –, applies to all doubly symmetric games. This means that in such games the dynamic process will always result in a 'local maximum' or 'local efficient' strategy.[21] From this it follows that in such games any local efficient strategy – which is itself already equivalent to being an ESS – is equivalent with asymptotic stability in the replicator dynamics.[22]

Fact 2 shows that a separating Nash equilibrium – i.e., a signalling equilibrium that according to Lewis is a potential linguistic convention – , will evolve in our evolutionarily language games (almost) by *necessity*.[23] The particular one that will evolve depends solely on the initial distribution of states and strategies (languages). With Skyrms (1996) we can conclude 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 conventional equilibria.

5 Reliability and costly signalling

Until now we have assumed that conventional languages are used only when the preferences of the agents involved are aligned. But, of course, we use natural language also if this pre-condition is (known) not (to be) met. As we have seen in section 2.2, however, in that case the sender (might) have an incentive to lie and/or mislead and the receiver has no incentive to trust what the sender claims. But even in these situations, agents – human or animal – sometimes send messages to each other, even if the preferences are less harmonically aligned.[24] Why would they do that? In particular, how could it be that natural language could be used for cooperative honest communication even in these unfavorable circumstances?

Perhaps the first answer that comes to mind involves *reputation* and an element of *reciprocity*. These notions are standardly captured in terms of the theory of *repeated* games (e.g.

[20] Our symmetric language games are *doubly* symmetric because for all $L_i, L_j, \mathcal{U}(L_i, L_j) = \mathcal{U}(L_j, L_i)$.

[21] Weibull also explains why this doesn't mean that in such games we will always reach the 'global' (or Pareto) optimal solution. As we will see in section 6, extra assumptions have to be made to guarantee this. Of course, Fisher's theorem holds in our games only because we made some idealizations, e.g. a simple form of reproduction (or learning) and perfect cooperation.

[22] This result generalizes to the evolutionarily stable set concept used by Wärneryd (1993).

[23] It is not an *absolute* necessity: if we start with two equally probable states and two messages, the mixture of strategies where all 16 possible languages are equally distributed is a stable state as well. In independent research, the (almost) necessity of emerging message-meaning relations is demonstrated also in *simulation*-models as those of Hurford (1989) and Oliphant (1996).

[24] To my surprise, Skyrms (manuscript) shows that some kind of information exchange is possible in evolutionary cheap talk games even in bargaining games where preferences are not aligned.

Axelrod & Hamilton, 1981).[25] The standard answer to our problem how communication can take place if the preferences are not perfectly aligned both in *economics* (starting with Spence, 1973) and in *biology* (Zahavi, 1975; Grafen, 1990; Hurd, 1995) doesn't make use of such repeated games. Instead, it is assumed that reliable communication is also possible in these circumstances, if we assume that signals can be too *costly* to fake.[26] The utility function of the sender takes no longer only the benefit of the receiver's action for a particular type of sender into account, but also the cost of sending the message. The aim of this section is to show that this standard solution in biology and economics can, in fact, be thought of as being very close to our intuitive solution involving reputation.

We will assume that the sender's utility function U_s can be decomposed in a *benefit function*, B_s and a *cost-function*, C. Consider now a two-type two-action game with the following benefit table.

two-type, two-action:

	a_H	a_L
t_H	$1,1$	$0,0$
t_L	$1,0$	$0,1$

In this game, the informed player (the sender) prefers, irrespective of her type, column player to choose a_H while column player wants to play a_H if and only if the sender is of type t_H. For a separating equilibrium to exist, individuals of type t_L must not benefit by adopting the signal typical of individuals of type t_H, even if they would elicit a more favorable response by doing so. Hurd (1995) shows that when we assume that the cost of sending a message can depend on the sender's type, an appealing separating equilibrium exists. Assume that the cost of message m saying that the sender is of type t_H is denoted by $C(t_i, m)$ for individuals of type i and that sending ϵ is costless for both types of individuals. Provided that $C(t_L, m) > 1 > C(t_H, m)$, the cost of sending m will outweigh the benefit of its production for individuals of type t_L, but not for individuals of type t_H, so that the following separating equilibrium exists: individuals of type t_H send message m, while individuals of type t_L send ϵ. Notice that on Hurd's characterization, in the equilibrium play of the game it is possible that not only t_L sends a costless message, but that the high type individual t_H does so as well![27] This suggests that the theory of costly signalling can be used to account for *honest* communication between humans who make use of a *conventional* language with *cost-free* messages. Moreover, an *evolutionary* argument shows that Hurd's characterization with cost-free messages sent in equilibrium is actually the most plausible one.[28] The only thing that really matters is that the cost of sending a deceiving message is higher than its potential benefit (so that they are sent only by individuals who deviate from equilibrium play). How can we guarantee this to be possible?

In the example discussed in this section, as in the examples discussed in the economic and biological literature, it is advantageous pretending to be better than one actually is.

[25] Gintis (2000) argues that such an explanation of cooperative behavior fails to predict cooperation when a group is threatened with extinction. He argues (with many others) that, instead, we should assume a form of *correlation* between individuals playing alike strategies to explain the evolution of cooperative behavior. Correlation can also be of help to resolve a worry one of the anonymous reviewers has: punishment itself can be thought of as being altruistic. We will come back to correlation in section 6.

[26] Also Asher et al. (2001) propose an analysis of Grice's quality maxim in terms of (a somewhat unusual version of) costly signalling. They don't relate it, however, with the standard literature in economics and biology.

[27] See Hurd (1995) for a more general characterization. This characterization differs from the one given by Grafen (1990) – which seems to be the one Zahavi (1975) had in mind –, according to which certain messages *cannot* be cost-free.

[28] Our game above not only has a separating equilibrium, but also a *pooling* one in which both types of individuals send ϵ and the receiver performs a_L. As it turns out, this pooling equilibrium cannot be evolutionarily stable if $C(t_H, m) < 1$. The same holds for separating equilibria where $C(t_H, m) > 0$.

This is crucially based on the assumption that messages are *not* (immediately) *verifiable*. This assumption opens the possibility that low-quality individuals could try to masquerade themselves as being of a high quality. And this assumption makes sense: if all messages could immediately be verified, the game being played is one of *complete information* in which it makes no sense to send messages about one's type (i.e. private information) at all. However, the assumption that messages are completely unverifiable is for many applications unnatural as well: an individual can sometimes be unmasked as a liar, and she can be punished for it. Thus, making a statement can be *costly*: one can be punished (perhaps in terms of reputation) when one has claimed to be better than one actually is.[29,30] If this punishment is severe enough, even a small probability of getting unmasked can already provide a strong enough incentive not to lie.[31]

The above sketched analysis of truthful human communication suggests that although natural language expressions are cheap in production, the theory of costly signalling can still be used to account for communicative behavior between humans. With Lachmann et al (manuscript) I take this to be an important insight: it suggests a way to overcome the limitations of both cheap talk signalling and the adoption of the cooperative assumption by Grice and Lewis. By assuming that sending signals can be costly, we can account for successful communication even if the preferences of the agents involved do not seem to be well aligned. Perhaps the most appealing way to think of Hurd's result is that it explains why in more situations the agent's preferences are aligned than it appears at first sight such that the possibility of communication is the rule, rather than the exception.[32]

[29]This way of looking at costs was brought to the author's attention by Carl Bergstrom (p.c.) and it's this way which brings us close to the conception of reciprocity.

[30]Of course, we don't need the theory of costly signalling to explain why no individual would say that she is worse than she actually is. Lying is not just not truly revealing one's type, but also doing this in such a way that it is (potentially) in one's own advantage.

[31]Lachmann et al (manuscript) argue – correctly we think – that the fact that the signalling costs are imposed *socially* by the receiver has two important consequences. First, the signaller now doesn't pay the costs associated with the signal level that she chose but rather with the signal level that the receiver *thinks* that she chose. As a consequence, in conventional signalling systems there will be selection for precise and accurate signals, in order to reduce costly errors. Second, in contrast to cases where costs are sender's responsibility, receivers have no incentive to reduce signal costs in the case we consider. As a consequence, the destabilizing pressure of selection for reduced signal costs will not be a threat to signalling systems in which cost is imposed by the signal receiver.

[32]Even if we assume that agents make use of signals with a pre-existing meaning and always tell the truth, this doesn't guarantee that language cannot be used to mislead one's audience. Take a familiar Gricean example. If an agent answers the question where John is by saying *John is somewhere in the South of France*, one might conclude that the agent doesn't know exactly where John is (see section 6.2 for the reason why) or that she doesn't think the exact place is relevant. However, it might be that she *does* know the exact place and knows that this is relevant, but just doesn't want to share this knowledge with the questioner. It all depends on the sender strategy taken, and this, in turn, depends on in how far the preferences of speaker and hearer are aligned. Look at the two-type-two-action game of this section again, assume that the expected utility for r to perform a_H is higher than that of a_L, and suppose that we demand truth: $t \in [[S(t)]]$. In that case, the rational message for a high-type individual to send is one that conventionally expresses $\{t_H\}$, while a low-type individual has an incentive to send a message with meaning $\{t_H, t_L\}$. If the receiver is naive he will choose a_H after hearing the signal that expresses $\{t_H, t_L\}$, because a_H has the highest expected utility. A receiver who knows the sender's strategy S, however, will realize that the proposition $\{t_H, t_L\}$ is only sent by a low type individual t_L, i.e., $S^{-1}(\{t_H, t_L\}) = \{t_L\}$, and thus will perform action a_L.

Obviously, when a hearer knows the sender-strategy being used by a speaker, deception is impossible. However, just as the uniqueness solution for coordination signalling problems, this is an unreasonably strong requirement to assume if it had to be determined anew for every separate conversational situation. Things would be much easier if for messages with a completely specified conventional meaning we can be assured that $[[m]] = S_m$, if S is the sender's strategy used in the particular conversation at hand. Without going into detail, we would like to suggest that this is again guaranteed by high costs of messages sent by individuals who deviate from equilibrium play, just like in the main text of this section.

6 The efficient use of language

Until now we have discussed how an expression m of the language used could come to have (and maintain) its conventional meaning $[[m]]$. This doesn't mean, however, that if a speaker uses m she just wants to inform the receiver that $[[m]]$ is the case. It is well established that a speaker normally wants to communicate more by the use of a sentence than just its conventional meaning. Sometimes this is the case because the conventional meaning of an expression *underspecifies* its actual truth-conditional interpretation; at other times the speaker *implicates* more by the use of a sentence than its truth-conditional conventional meaning. It is standard to assume that both ways of enriching conventional meaning are possible because we assume that the speaker conforms to Grice's (1967) *maxims of conversation*: she speaks the truth (*quality*), the whole truth (*quantity*), though only the relevant part of it (*relevance*), and does so in a clear and efficient way (*manner*). Grice argued that because the speakers are taken to obey these maxims, a sentence can give rise to *conversational implicatures*: things that can be inferred from an utterance that are not conditions for the truth of the utterance. Above, we discussed already the maxim of quality, which has a somewhat special status. Grice argues that the implicatures generated by the other maxims come in two sorts: *particularized ones*, where the implicature is generated by features of the context; and *generalized* ones, where (loosely speaking) implicatures are seen as default rules possibly overridden by contextual features. There exist general agreement that both kinds of implicatures exist, but the classification of the various implicatures remains controversial within pragmatics. Whereas relevance theorists (e.g. Sperber & Wilson, 1986) tend to think that implicatures depend predominantly on features of the particular context, Levinson (2000), for example, takes generalized implicatures to be the rule rather than the exception. Similar controversies can be observed on the issue of how to resolve underspecified meanings: whereas Parikh (1991, 2001) argues optimistically that indeterminacy in natural language can be solved easily in many cases through the existence of a unique (Pareto-Nash) solution of the coordination problem of how to resolve the underspecification, proponents of centering theory (Grosz et al. 1995), for example, argue that pronoun resolution is, or needs to be, governed by structural (default) rules.

Except for the maxim of quality, Horn (1984), Levinson (2000), and others argue that the Gricean maxims can be reduced to two general principles: The I-principle which tells the hearer to interpret a sentence in its most likely or stereotypical way, and the Q-principle which demands the speaker to give as much (relevant) information as possible. In this section it will be argued that two general pragmatic rules which closely correspond with these two principles can be given an evolutionary motivation which suggests that 'on the spot' reasoning need not play the overloaded role in natural language interpretation as is sometimes assumed.

6.1 Iconicity in Natural Languages

In section 3.1 we have seen that Lewis (1969) proposes to explain the semantic/conventional meaning of expressions in terms of separating equilibria of signalling games. However, we also saw that simple costless signalling games have many such equilibria. Lewis assumed that each of these equilibria are equally good and thus that it is completely *arbitrary* which one will be chosen as a convention. In section 4 we have seen that all separating equilibria satisfy the ESS condition and that which one will in the end emerge is a matter of chance and depends only on the initial distribution of states and strategies (languages). Although natural at the level of individual words and the objects they refer to, at a higher organizational level the assumption of pure arbitrariness or chance can hardly be sustained. It

cannot explain why conventions that enhance *efficient* communication are more likely than others that don't.

Consider a typical case of communication where two meanings t_1 and t_2 can be expressed by two linguistic messages m_1 and m_2. We have here a case of *underspecification*: the same message can receive two different interpretations. In principle this gives rise to two possible codings: $\{\langle t_1, m_1\rangle, \langle t_2, m_2\rangle\}$ and $\{\langle t_1, m_2\rangle, \langle t_2, m_1\rangle\}$. In many communicative situations, however, the underspecification does not really exist, and is resolved due to the general pragmatic principle – referred to as the pragmatic *iconicity principle* – that a lighter (heavier) form will be interpreted by a more (less) salient, or stereotypical, meaning: (i) It is a general defeasible principle, for instance, in centering theory (Grosz et al, 1995) that if a certain object/expression is referred to a pronoun, another more salient object/expression should be referred to by a pronoun too; (ii) Levinson (2000) seeks to reduce Chomsky's B and C principles of the binding theory to pragmatics maxims. In particular, disjoint reference of lexical noun phrases throughout the sentence is explained by pointing to the possibility of the use of a lighter expression, viz. an anaphor or pronoun; (iii) The preference for stereotypical interpretations (Atlas & Levinson, 1981); (iv) and perhaps most obviously, Horn's (1984) division of pragmatic labor according to which an (un)marked expression (morphologically complex and less lexicalized) typically gets an (un)marked meaning (cf. *John made the car stop* versus *John stopped the car*).

Horn (1984), Levinson (2000), Parikh (1991, 2001) and Blutner (2000) correctly suggest, that because this generalized pragmatic iconicity principle allows us to use language in an *efficient* way, it is *not* an *arbitrary convention* among language users. There is no alternative rule which would do equally well for the same class of interactions if people generally conformed to this alternative. This can be seen most simply if we think of languages that are separating equilibria in our language game as *coding systems* of meanings distributed with respect to a particular probability function.[33] This suggests that the rule should follow from more general economic principles. Indeed, Parikh gives a game theoretical analysis of why this principle of iconicity is observed. However, he treats it as a *particularized conversational implicature*. Here we want to argue that it should rather be seen as a *generalized* default rule.[34]

6.1.1 Underspecification and Pragmatic interpretation rules

In section 2.2 we saw that in cheap talk games meaningful communication is possible only in so far as the preferences of the participants coincide. But in section 5 we showed that by making use of costly messages we can overcome this limitation. It is standardly assumed that this is the only reason why costs of messages are taken into account: to turn games in which the preferences are not aligned to ones where they are. We have suggested that in this way we can account for Grice's maxim of quality. In this section we will see, however, that costly messages can also be used to account for another purpose and give a motivation for the pragmatic iconicity principle.

To be able to do so, we should allow for *underspecification* or *context dependence*. In

[33]Suppose that S is the sender strategy of a signalling system and P the probability function over states. A general fact of Shannon's (1948) information theory is that if S is an optimal coding of the meanings, it will be the case that if $P(t) < P(t') < P(t'')$ then $l(S(t)) \geq l(S(t')) \geq l(S(t''))$, where $l(S(t))$ is the length of expression $S(t)$.

[34]See van Rooy (to appear, though dating back to 2001) for a more extensive argument against Parikh's analysis. There it is also argued that Horn's division of pragmatic labor follows from an evolutionary stance on signalling games once we have underspecification. In this paper we go a step further, and also give an evolutionary motivation for why underspecification itself is so useful.

different contexts, the same message can receive a different interpretation.[35] In our description of signalling games so far it is not really possible to represent a conventional language with underspecified meanings that are resolved by context. The best thing we could do is to represent underspecification as real *ambiguity*: sender strategy S is a function that assigns the same message to different states, while receiver strategy R is a *mixed* strategy assigning to certain messages a non-trivial probability distribution over the states. Such a sender-receiver strategy combination can never be evolutionarily stable (cf. Wärneryd, 1993): one can show that a group of individuals using a mutant language without such ambiguity has no problem invading and taking over a population of ambiguous language users (if there exists an unused message).

To account for underspecification, we have to enrich our models and take *contexts* into account. For the purpose of this section we can think of a context as a probability distribution over the state space T. For simplicity (but without loss of generality) we assume that $T = \{t, t'\}$ and $M = \{m, m'\}$. Communication takes place in two kinds of contexts: in one context where $P(t) = 0.9$ (and thus $P(t') = 0.1$) and in one where $P(t) = 0.1$. We assume that both contexts are equally likely. Call the first context ρ_1 and the second ρ_2. We will assume that it is common knowledge among the conversational partners in which context they are, but only the sender knows in each context in which state she is. The messages don't have a pre-existing meaning, but differ in terms of (production) costs: we assume that $C(m) < C(m')$. However, 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 section 5, we assume that the cost of sending a message can never exceed the benefit of communication. To assure this, we will take the sender's utility function to be decomposable into a benefit and a cost function, $U_s(t_i, m_j, t_k) = B_s(t_i, t_k) - C(m_j)$, with $C(m) = 0, C(m') = \frac{1}{3}$, and adopt the following benefit function:

$$\begin{aligned} B_s(t_i, t_k) &= 1, \text{ if } t_k = t_i \\ &= 0 \text{ otherwise} \end{aligned}$$

A sender strategy is now a function mapping a state *and a context* to a message, while a receiver strategy is now a function from message-context pairs to states. This game has of course two separating equilibria, call them L and L', with no underspecification: L gives rise to the mapping $\{\langle t, m\rangle, \langle t', m'\rangle\}$ in both contexts, while L' to $\{\langle t, m'\rangle, \langle t', m\rangle\}$. These languages are also evolutionarily stable in the sense of being an ESS. However, our new game also allows for languages with underspecification to be evolutionarily stable. Given that it is common knowledge between sender and receiver in which context they are, the only requirement is that they give rise to a *separating* equilibrium *in each context*. We can distinguish two such underspecified languages: (i) the Horn language L_H with the mappings $\{\langle t, m\rangle, \langle t', m'\rangle\}$ and $\{\langle t, m'\rangle, \langle t', m\rangle\}$ in contexts ρ_1 and ρ_2, respectively; and (ii) the anti-Horn language L_{AH} where the two mappings are used in the other contexts. It is easily seen that both languages are evolutionarily stable, because also $\langle L_H, L_H\rangle$ and $\langle L_{AH}, L_{AH}\rangle$ are *strict* Nash equilibria. Both languages do better against themselves than against the other, or against L or L'.

Our above discussion shows that underspecification is possible. However, we want to explain something more: why is underspecification useful, and why is the underspecified Horn language L_H which incorporates the pragmatic iconicity principle more natural to emerge than the underspecified Anti-Horn language L_{AH}?

[35] In lexical semantic terms, we have to account for the fact that *homonymy* and *polysemy* are natural in languages.

As it turns out, the problem we encountered is a well-known one in game theory: how to select among a number of *strict* Nash equilibria the one that has the highest expected utility, i.e., is (in our games) *Pareto optimal*? In our language game described above we had four strict equilibria: $\langle L, L\rangle, \langle L', L'\rangle, \langle L_H, L_H\rangle$, and $\langle L_{AH}, L_{AH}\rangle$. These equilibria correspond to our four evolutionarily stable languages L, L', L_H, and L_{AH}, respectively. A simple calculation shows that the Horn language is the one with the highest expected utility.[36] Thus, if we can find a natural explanation of why our evolutionary dynamics tends to select such optimal equilibria, we have provided a naturalistic explanation for why (i) languages make use of underspecification, and (ii) respect the iconicity principle.

6.1.2 Correlated and Stochastic evolution

In van Rooy (in press), two (relatively) standard explanations of why Pareto optimal solutions (in coordination games) tend to evolve are discussed. According to both, we should give up an assumption behind the stability concepts used so far.

According to the first explanation (e.g. Skyrms, 1994, 1996) we give up the assumption that individuals *pair randomly* with other individuals in the population. Random pairing is assumed in the calculation of the expected utility of a language. The probability with which individuals using language L_i interact with individuals using L_j depends simply on the proportion of individuals using the latter language: $\mathcal{EU}(L_i) = \sum_j P(L_j) \times \mathcal{U}(L_i, L_j)$. This expected utility was used both to determine the ESS concept and to state the replicator dynamics. By giving up random pairing (a well-known strategy taken in biology to account for kin-selection, and in cultural evolution to account for clustering), we have to postulate the existence of an additional function which determines the likelihood that an individual playing L_i encounters an individual playing L_j, $\pi(L_j/L_i)$, such that $\sum_j \pi(L_j/L_i) = 1$. What counts then is the following expected utility: $\mathcal{EU}_\pi(L_i) = \sum_j \pi(L_j/L_i) \times \mathcal{U}(L_i, L_j)$. The other definitions used in the dynamic system follow the standard definitions in replicator dynamics.

Although this generalization seems to be minor, it can have significant effects on the resulting stable states. Assume a form of *correlation*: a tendency of individuals to interact more with other individuals playing the same strategy (i.e., using the same language). Formally, positive correlation comes down to the condition that for any language L_i, $\pi(L_i/L_i) > P(L_i)$ (see Skyrms, 1994). In the extreme case, i.e. $\pi(L_i/L_i) = 1$, the only stable state in the replicator dynamics is the one which has the highest expected utility in self-interaction. In our case the Pareto optimal language L_H is selected and we have an evolutionary explanation for the existence of underspecification and the use of iconicity.[37]

For our evolutionary language game there is another, and perhaps more natural, possibility to ensure the emergence of Pareto efficient languages. It is to give up the assumption that the transition from one generation to the next in the dynamic model is completely determined by the distribution of strategies played in a population and their expected utilities.

[36] Taking $C(m') = \frac{1}{3} = c$, the expected utility of L in self-interaction can be determined as follows:

$$\begin{array}{rcl} \mathcal{U}(L,L) & = & [\frac{1}{2} \times \mathcal{U}(\rho_1, L, L)] + [\frac{1}{2} \times \mathcal{U}(\rho_2, L, L)] \\ & = & [\frac{1}{2} \times (0.9 + 0.1(1-c)] + [\frac{1}{2} \times (0.9(1-c) + 0.1)] \\ & = & [\frac{1}{2} \times (\frac{27}{30} + \frac{2}{30})] + [\frac{1}{2} \times (\frac{18}{30} + \frac{3}{30})] \\ & = & \frac{29}{60} + \frac{21}{60} = \frac{25}{30} \end{array}$$

Similar calculations show that $\mathcal{U}(L', L') = \frac{25}{30}$, $\mathcal{U}(L_H, L_H) = \frac{29}{30}$, and $\mathcal{U}(L_{AH}, L_{AH}) = \frac{21}{30}$.

[37] The assumption that $\pi(L_i/L_i) = 1$ is unnecessary strong. Computer simulations suggest that much milder forms of correlation have already the desired effect, albeit at a larger time scale. See `http://signalgame.blehq.org` for an implementation (mainly due to Wouter Koolen) of the evolutionary language game with correlation and mutation and play with it yourself!

We can assume that the transition is (mildly) *stochastic* in nature.[38] As shown by Kandori, Mailath & Rob (1993) and Young (1993), this results in the selection of the so-called *risk-dominant* strict Nash equilibria in the (very) long term.[39,40] In general, a risk-dominant equilibrium need not be Pareto efficient, but in cooperative games the two concepts coincide.

A natural way to allow for stochastic adjustment in our evolutionary language game is to give up the assumption that an individual simply adopts the strategy from its parent with probability 1. Giving up this assumption makes sense: the inheritance of language is imperfect, possibly due to non-optimal *learning*.[41]

It is not obvious which of the proposals to motivate the attraction of Pareto optimal solutions is more plausible to assume for natural languages. In fact, both factors (clustering and innovation) seem to play a role in the evolution and change of languages.

6.2 Exhaustive interpretation

According to Grice's (1967) maxim of Quantity, or Horn's (1984) and Levinson's (2000) Q-principle, speakers should give as much (relevant) information as possible. In this section we will give a motivation for that, and suggest a way to explain the convention according to which the effort required for this optimal transfer of information is divided between speaker and hearer such that the speaker doesn't have to be fully explicit and the receiver interprets answers *exhaustively*.

Let us assume that the *relevant* information is the information needed for the receiver to resolve his decision problem. Suppose the receiver r is commonly known to be a Bayesian utility maximizer and he asks question $\mathcal{Q}$ because this question 'corresponds' closely to his decision problem. What is then the best answer-strategy S to give relevant information? Let us look at some candidates.

According to strategy S_1, the speaker gives as much relevant information as she can. Assume that $K(t)$ denotes the set of states that the sender thinks are possible in t. Assume, furthermore, that question $\mathcal{Q}$ gives rise to (or means) a partition of T: Q. Thus, each element q of Q is also a set of states. We will refer both to the interrogative sentence and to the induced partition as a question. Then we define $Q_{K(t)}$ to be $\{q \in Q | q \cap K(t) \neq \emptyset\}$, i.e. the elements of Q which the sender takes to be possible. Then S_1 is the strategy that gives in every state t the following proposition: $\bigcup Q_{K(t)}$, i.e., the union of the elements of $Q_{K(t)}$. Suppose, for instance, that Q denotes the question corresponding to 'Who came?', i.e., 'Which individuals have property P?', and that $K(t) = \{t, t'\}$ such that in t (only) John came, and in t', John and Mary. The message that then expresses proposition $\bigcup Q_{K(t)}$ is

[38] Very recently, Gerhard Jäger started to make use of this assumption as well in his manuscript 'Evolutionary Game Theory and Typology: A Case Study' to account for the restricted distribution of case-marking systems among natural languages.

[39] Consider Rousseau's famous Stag Hunt game as described by Lewis (1969); a simple two-player symmetric game with two strict equilibria: both hunting Stag or both hunting Rabbit. The first is Pareto optimal because it gives to both a utility of, let us say, 6, while the second only one of 4. However, assume that if one hunts Stag but the other Rabbit, the payoff is (4,0) in 'favor' of the Rabbit-hunter. In this case, the non-Pareto equilibrium where both are hunting Rabbit is called risk-dominant, because the 'resistance' of $\langle R, R\rangle$ against $\langle S, S\rangle$ is greater than the other way around. The reason for this, intuitively, is that if one player is equally likely to play either strategy, the expected utility of hunting Rabbit for the other is optimal.

[40] As Gerhard Jäger reminded me of, this is only proved for 2-by-2 games, although I assume it also holds for the more general case. Fortunately, Jäger also assures me that simulation models suggest that the statement holds more generally in our doubly symmetric games.

[41] Lightfoot (1991) and others show that this might have interesting consequences for language change. In a number of papers by Nowak and colleagues (e.g. Komarova et al., 2001), evolutionary dynamics is used to study the requirements on the language acquistition device in order for languages to remain stable. According to a lot of historical linguists (e.g. Croft, 2001), however, the influence of innovative language use by adults should not be underestimated.

'John, perhaps Mary, and nobody else'. If we assume that the answerer knows exactly who came, i.e., is known to be fully *competent* about the question-predicate P, the proposition expressed is $\lambda t'[P(t') = P(t)]$.

According to strategy S_2, the speaker gives the set of individuals of whom she is certain that they satisfy the question-predicate. Thus, for the question *Who has property P?*, the answer is going to be $\lambda t'[KP(t) \subseteq P(t')]$, where $KP(t)$ is the set $\{d \in D | K(t) \models P(d)\}$ and $K(t) \models P(d)$ iff $\forall t' \in K(t) : d \in P(t')$. In the example discussed for strategy S_1, the answerer would now use a message like '(At least) John came'. Notice that if the answerer is known to be competent about the extension of P, the answer reduces to $\lambda t'[P(t) \subseteq P(t')]$.

These answer-strategies closely correspond with some well-known analyses of questions in the semantic literature: on the assumption of competence, S_1 gives rise to Groenendijk & Stokhof's (1984) partition semantics: $\{S_1(t)|t \in T\} = \{\lambda t'[P(t') = P(t)]|t \in T\}$, while S_2 gives rise to $\{S_2(t)|t \in T\} = \{\lambda t'[P(t) \subseteq P(t')]|t \in T\} = \{\bigcap\{\lambda t'[d \in P(t')]|d \in P(t)\}|t \in T\}$ which corresponds to Karttunen's (1977) semantics for questions.

In order to investigate which of S_1, S_2, or some other strategies can be part of a Nash equilibrium together with receiver strategy R_B which implements a Bayesian rational agent, we have to assume that, in equilibrium, the receiver knows S. Thus he is not going to update his belief with $[[S(t)]]$, but rather with $S_t = \{t' \in T | S(t') = S(t)\}$.

A speaker who uses S_1 would give in each state at least as much (relevant) information as speakers using the alternative strategies. In particular, for each state t (where P has a non-empty extension), $S_{1,t} \subseteq S_{2,t}$. This is obvious for the propositions that would be given in state t on the assumption of full competence: $S_1 : \lambda t'[P(t') = P(t)]$; $S_2 : \lambda t'[P(t) \subseteq P(t')]$, but the same is true if we don't make our assumption of competence. A well-known fact of decision theory (Blackwell, 1953) states that an agent with an information structure, or possibility operator, K, is able to make at least as good decisions as an agent with possibility operator K' iff for each $t \in T : K(t) \subseteq K(t')$. But this means that if our questioner is Bayesian rational and is going to believe what the answerer tells him, S_1 is the for him preferred answer-strategy. On an assumption of perfect cooperation, this is also true for the answerer himself. We can conclude that if R_B is the strategy adopted by a perfect Bayesian, $\langle S_1, R_B \rangle$ is the only *Nash equilibrium* of the game induced by question $\mathcal{Q}$. In fact, it is (on average) *strictly* better than S_2 and other alternatives, which means that $\langle S_1, R_B \rangle$ is the only ESS on the same assumptions. Thus, in cooperative games it is optimal to obey the Gricean maxim to give as much relevant information as possible, i.e. give an answer with meaning $\bigcup Q_{K(t)}$.

Suppose that the question under discussion, $\mathcal{Q}$, is *Who has property P?*. How should the answer be *coded*? Given that it is only propositions involving question-predicate P that counts, the optimal answer $\bigcup Q_{K(t)}$ given in state t equals the intersection of the following propositions (where D is the set of individuals, $\Diamond P(d)$ means that the speaker thinks it is possible that d has property P, and $\bar{P}$ denotes the complement of P):

$$W = \bigcap_{d \in D} \{[[P(d)]] : K(t) \models P(d)\}; \qquad X = \bigcap_{d \in D} \{[[\Diamond P(d)]] : K(t) \cap [[P(d)]] \neq \emptyset\};$$

$$Y = \bigcap_{d \in D} \{[[\Diamond \bar{P}(d)]] : K(t) \cap [[\bar{P}(d)]] \neq \emptyset\}; \qquad Z = \bigcap_{d \in D} \{[[\bar{P}(d)]] : K(t) \models \bar{P}(d)\}.$$

Now suppose that the answerer is mutually known to be fully competent about predicate P. That is, she knows of each individual whether it has property P or property $\bar{P}$. In that case $\bigcup Q_{K(t)} = W \cap Z$, i.e., the proposition that states for each individual whether it has property P or property $\bar{P}$. We have seen above that the combination $\langle S_1, R_B \rangle$ is an equilibrium strategy as far as information transfer is concerned. But it is a somewhat unfair

equilibrium as well: the sender is (normally) required to give a very complex answer, while the receiver can lay back. As far as information transfer is concerned, however, our above reasoning doesn't force us to this equilibrium. If it is common knowledge between speaker and hearer that the latter will infer more from the use of a message than its conventional meaning, many other sender-receiver strategies give rise to the same optimal information transfer as well. For example, on our assumption of full competence again, the speaker doesn't have to explicitly state of each individual whether it satisfies question-predicate P or satisfies $\bar{P}$. If it is mutually known between the two that the sender only mentions the positive instances, i.e., uses strategy S_2, she can just express the above proposition W, leaving proposition Z left for the hearer to infer. This would obviously be favorable to the sender, but is a natural consequence of evolution, only if the extra effort transferred to the hearer is relatively small. This is possible, if the task left to the hearer – i.e., to infer from a message with conventional meaning W to the above proposition $\bigcup Q_{K(t)} = W \cap Z$ –, can be captured by a *simple* but still *general* interpretation mechanism. As it turns out, this is, in fact, the case. Assuming that the hearer receives message m as answer to the question *Who has property P?*, he can interpret it as follows (where $t' <_P t$ iff t is just like t' except that P has a smaller extension in t' than in t).

$$R^P_{exh}(m) \quad = \quad \{t \in [[m]] | \neg \exists t' \in [[m]] : t' <_P t\}$$

This interpretation strategy of answers is known as *predicate circumscription* (of m with respect to P) in Artificial Intelligence (McCarty, 1980). In linguistics it is known as the *exhaustive interpretation* of an answer (Groenendijk & Stokhof, 1984). In both fields it has become clear that such an interpretation method can account for many inferences concerning what the speaker meant, but did not say: what is implicitly conveyed by message m is according to this mechanism everything that follows from $R^P_{exh}(m)$ but not yet from $[[m]]$. In vanRooy & Schulz (manuscript) it is shown that it can account for many implicatures discussed in the semantic/pragmatic literature, most importantly the ones that are usually accounted for in terms of Grice's (first sub)maxim of quantity.[42] Thus, it appears that the *general* strategy to interpret answers exhaustively is *simple* enough to make sense from an economical/evolutionary perspective, and can account for many things we infer from the use of a sentence on top of its conventional meaning.[43] We don't need as much 'on the spot' reasoning involving a strong notion of rationality to account for these inferences as is sometimes assumed: we only have to apply the interpretation rule and do not have to reason *that* we should apply it.

7 Conclusion and Outlook

In this paper we have discussed conventional and non-conventional interpretation strategies, and contrasted two ways of answering the question of why these strategies are chosen: in terms of *standard*, or *rational*, and *evolutionary* game theory. We have argued that natural language interpretation doesn't need to rely as much on principles of rationality and assumptions of common knowledge as is sometimes assumed, and that taking an evolutionary

[42] It is also shown how some apparent counterexamples to the use of this mode of interpretation to account for implicatures can be overcome when we bring it together with some modern developments in semantics/pragmatics.

[43] Of course, a combined speaker-hearer strategy such that the sender only gives the negative instances and the receiver adopts a strategy just like the one in the main text except that the order on T is based on $\bar{P}$ instead of on P would do equally well. This would, however, be more costly if there are more negative than positive instances of predicate P, which is normally, though not always, the case. Therefore, the rule which only gives the positive instances is from an economical point of view more natural.

stance solves some problems which rational game theory leaves unresolved. First, building on work of economists and biologists, it was shown that Lewis's (1969) rationalistic analysis of *semantic* conventions could be given a natural evolutionary alternative. In the second part of the paper we suggested the same for some general *pragmatic* principles. Although we agree with Grice and others that the principles of *truthfulness*, *iconicity*, and *exhaustive interpretation* are not arbitrary and should be based on extra-linguistic economic principles, this doesn't necessarily mean that agents observe these principles because they are fully rational and come with a lot of common knowledge in each particular conversational situation. We argued that also the process of pragmatic interpretation might to a large extent be *rule*-governed, and motivated some of these rules in terms of evolutionary game theory. This doesn't mean that pragmatics makes no use of 'on the spot' reasoning, but its task is perhaps not as important as is sometimes assumed. A major task for the future is to delve more deeply into the question of how the interpretation task should be divided between linguistic rules and on the spot reasoning.

A further task is to extend our analysis of signalling. The strategies in the signalling games we have discussed until now have obvious limitations and can hardly be called languages. Most obviously, we assumed that messages are just unstructured wholes, although it is standardly assumed that the meanings of natural language sentences are determined *compositionally* in terms of their parts.[44] It is this compositionality that allows speakers and hearers to continually produce and understand new sentences, sentences that have never before been uttered or heard by that speaker or hearer.

How did humans come from our simple signalling systems to complicated languages and what was their incentive to do so? Despite the fact that the Académie des Science declared around 1850 any proposal to answer these questions as pure speculation and thus unscientific, very recently we saw a renewed interest from authors with scientific aspirations to address exactly these issues. In the literature, two kinds of answers are given to the above two questions.

The first answer to the question *why* compositionality arose that comes to mind is that compositionality allows speakers to communicate about a greater number of situations with obvious adaptive advantages. This answer seems to presuppose that *memory* is very limited. Perhaps more limited than it actually is. The second, and most standard, answer is that *learnability* rather than communicative success is the stimulating factor (e.g. Kirby & Hurford, 2001). If we want to talk about lots of different (aspects of) situations, a conventional language needs to distinguish many messages. In order for such a language to remain stable over generations, precedent or imitation are unnatural explanatory mechanisms: there are just too many message-meaning combinations to be learned. Once languages have a compositional structure, this problem disappears: we only need a small finite number of conventions which *can* plausibly be learned by children, but which together deductively entail a possibly infinite number of conventional associations between messages and meanings.

According to the, perhaps, standard answer to the question *how* compositionality arose, it is assumed that meaningful messages are really *words* and that different kinds of words – ones that denote objects ('nouns') and ones that denote actions ('verbs') – can be combined together to form (subject-predicate) sentences. This analysis assumes that both the set of states (our set T) and the set of messages (M) are already partitioned into objects and actions, and nouns and verbs, respectively. This approach is worked out in some detail by Nowak & Krakauer (1999), among others, and dubbed *synthetic* in Hurford (2000).

Wray (1998) – in the footsteps of the Quinean "radical translation" tradition – has objected to this approach. The meaningful messages in primitive communication systems

[44] For a fuller list of limitations, see Lewis (1969, pp. 160-61).

should not be thought of as words, but rather as whole *utterances* that describe particular kinds of *situations*. Compositional systems do not arise through the combination of meaningful words, but rather through the correlation between (i) features of meaningful messages and (ii) aspects of the situations that these utterances describe. Instead of assuming that the sets of states and messages are already partitioned, it is better to suppose that the states and messages themselves are structured and represented by something like vectors. Ideas of this kind are worked out (in terms of computer simulations) by Steels (2000), Kirby & Hurford (2001), and others, and such approaches are called *analytic* by Hurford (2000). This work is very appealing. It has not yet been given a theoretical underpinning within EGT. This would be very useful, because it would provide the experimental results an analytic justification. But this only means that we have something to look forward to!

References

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

[2] Axelrod, R. and W. Hamilton (1981), 'The evolution of cooperation', *Science*, **411**: 1390-1396.

[3] Blackwell, D. (1953), 'Equivalent comparisons of experiments', *Annals of Mathemathical Statistics*, **24**: 265-72.

[4] Blutner, R. (2000), 'Some aspects of Optimality in Natural Language Interpretation', *Journal of Semantics*, **17**: 189-216.

[5] Clark, H. (1996), *Using Language*, Cambridge University Press, Cambridge.

[6] Crawford, V. and J. Sobel (1982), 'Strategic information transmission', *Econometrica*, **50**: 1431-51.

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

[8] Dretske, F. (1981), *Knowledge and the Flow of Information*, MIT Press, Cambridge, Massachusetts.

[9] Fisher, R.A. (1930), *The Genetical Theory of Natural Selection*, Oxford University Press, Oxford.

[10] Grafen, A. (1990), 'Biological signals as handicaps', *Journal of Theoretical Biology*, **144**: 517-546.

[11] Gintis, H. (2000), *Game Theory Evolving*, Princeton University Press, Princeton.

[12] Grice, H.P. (1957), 'Meaning', *Philosophical Review*, **66**: 377-388.

[13] Grice, H. P. (1967), 'Logic and Conversation', typescript from the William James Lectures, Harvard University. Published in P. Grice (1989), *Studies in the Way of Words*, Harvard University Press, Cambridge Massachusetts, 22-40.

[14] Groenendijk, J. and M. Stokhof (1984), *Studies in the Semantics of Questions and the Pragmatics of Answers*, Ph.D. thesis, University of Amsterdam.

[15] Grosz, B., A. Joshi and S. Weinstein (1995), 'Centering: A framework for modeling the local coherence of discourse', *Computational Linguistics*, **21**: 203-226.

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

[17] Hurd, P. (1995), 'Communication in discrete action-response games', *Journal of Theoretical Biology*, **174**: 217-222.

[18] Hurford, J. (1989), 'Biological evolution of the saussurian sign as a component of the language acquisition device', *Lingua*, **77**: 187-222.

[19] Hurford, J. (2000), 'The Emergence of Syntax', In: C. Knight, M. Studdert-Kennedy and J. Hurford (eds.), *The Evolutionary Emergence of Language: Social function and the origins of linguistic form*, (editorial introduction to section on syntax), Cambridge University Press. pp. 219-230.

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

[21] Karttunen, L. (1977), 'Syntax and semantics of questions', *Linguistics and Philosophy*, **1**: 3–44.

[22] Kirby, S. and J. Hurford (2001), 'The Emergence of Linguistic Structure: an Overview of the Iterated Learning Model'. In: D. Parisi & A. Cangelosi (eds.), *Simulating the Evolution of Language*, Springer Verlag, Berlin.

[23] Komarova, N., P. Niyogi and M. Nowak (2001), 'The evolutionary dynamics of grammar acquisition', *Journal of Theoretical Biology*, **209**: 43-59.

[24] Lachmann, M., Sz. Szamado, and C. Bergstrom (manuscript), 'The peacock, the sparrow, and evolution of human language', Santa Fe Institute working paper nr. 00-12-74.

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

[26] Lewis, D. (1969), *Convention*, Harvard University Press, Cambridge, Massachusetts.

[27] Lightfoot, D.W. (1991), *How to Set Parameters: arguments from language change*, MIT Press, Cambridge, Massachusetts.

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

[29] Maynard-Smith, J. (1982), *Evolution and the Theory of Games*, Cambridge University Press, Cambridge.

[30] Maynard-Smith, J. and D. Harper (1995), 'Animal Signals: Models and Terminology', *Journal of Theoretical Biology*, **177**: 305-311.

[31] McCarthy, J. (1980), 'Circumscription - a form of non-monotonic reasoning', *Artificial Intelligence*, **13**: 27 - 39.

[32] Millikan, R.G. (1984), *Language, Thought, and Other Biological Categories*, MIT Press, Cambridge, Massachusetts.

[33] Nowak, M. and D. Krakauer (1999), 'The evolution of language', *Proc. Natl. Acad. Sci. U.S.A.*, **96**: 8028-8033.

[34] Oliphant, M. (1996), 'The dilemma of saussurean communication', *BioSystems*, **37**: 31-38.

[35] Osborne, M. and A. Rubinstein (1994), *A course in Game Theory*, MIT Press, Cambridge, Massachusetts.

[36] Parikh, P. (1991), 'Communication and Strategic Inference', *Linguistics and Philosophy*, **14**: 473-513.

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

[38] Rooy, R. van (2003), 'Quality and quantity of information exchange', *Journal of Logic, Language and Information*, **12**, 423-451.

[39] Rooy, R. van (in press), 'Signaling games select Horn strategies', *Linguistics and Philosophy*, to appear.

[40] Rooy, R. van and K. Schulz (manuscript), 'Pragmatic Meaning and Non-monotonic Reasoning: The Case of Exhaustive Interpretation', University of Amsterdam.

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

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

[43] Selten, R. (1980), 'A note on evolutionary stable strategies in asymmetric animal contests', *Journal of Theoretical Biology*, **84**: 93-101.

[44] Searle, J.R. (1969), *Speech acts*, Cambridge University Press, Cambridge.

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

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

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

[48] Skyrms, B. (manuscript), 'Signals, Evolution and the Explanatory Power of Transient Information', University of California, Irvine.

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

[50] Sperber, D. and D. Wilson (1986), *Relevance*, Harvard University Press, Cambridge.

[51] Steels, L. (2000), 'The Emergence of Grammar in Communicating Autonomous Robotic Agents', In: W. Horn, (ed.), *Proceedings of European Conference on Artificial Intelligence, ECAI2000, nr. CONF14*, Amsterdam: IOS Press, pp. 764-769.

[52] Strawson, P.F. (1964), 'Intention and convention in speech acts', *Philosophical Review*, **75**: 439-460.

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

[54] Tuyls, K. et al. (this volume), 'An evolutionary game theoretical perspective on learning in multi-agent systems', *Knowledge, Rationality and Action.*

[55] Vanderschraaf, P. (1995), 'Convention as correlated equilibrium', *Erkenntnis*, **42**: 65-87.

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

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

[58] Wray, A. (1998), 'Protolanguage as a holistic system for social interaction', *Language and Communication*, **19**: 47-67.

[59] Young, H.P. (1993), 'The evolution of conventions', *Econometrica*, **61**: 57-84.

[60] Zahavi, A. (1975), 'Mate selection – a selection for a handicap', *Journal of Theoretical Biology*, **53**: 205-214.