Evolution and the Explanation of Meaning*

Simon M. Huttegger†‡

Signaling games provide basic insights into some fundamental questions concerning the explanation of meaning. They can be analyzed in terms of rational choice theory and in terms of evolutionary game theory. It is argued that an evolutionary approach provides better explanations for the emergence of simple communication systems. To substantiate these arguments, I will look at models similar to those of Skyrms (2000) and Komarova and Niyogi (2004) and study their dynamical properties. My results will lend partial support to the thesis that evolution leads to communication. In general, states of partial communication may evolve with positive probability under standard evolutionary dynamics. However, unlike states of perfect communication, they are unstable relative to neutral drift.

1. Introduction. Signaling games model simple signaling interactions between a sender and a receiver. They emphasize the social aspects of language. Thus, they are able to provide basic insights into some fundamental questions concerning the explanation of how meaningful communication can emerge. The importance of societal relations for the emergence of communication was already pointed out by David Hume (1739), Jean-Jacques Rousseau (1755), and Adam Smith (1761). It was further elaborated in David Lewis's (1969) seminal work on conventions. One main

*Received June 2005; revised February 2007.

†To contact the author, please write to: Konrad Lorenz Institute for Evolution and Cognition Research, Adolf Lorenz Gasse 2, A-3422 Altenberg, Austria; e-mail: simon.huttegger@kli.ac.at.

‡I would like to thank Brian Skyrms for his advice and encouragement. I am also grateful to Georg Dorn, Bill Harms, Josef Hofbauer, Gerhard Jäger, Natalia Komarova, Hannes Leitgeb, Don Saari, Kevin Zollman, two anonymous referees, and the editor of *Philosophy of Science* for many helpful suggestions and comments. Most of this paper was written during a stay at the Department of Logic and Philosophy of Science at the University of California at Irvine, whose members I wish to thank for their hospitality and kindness. My research was supported by the Institut für Wissenschaftstheorie of the Internationales Forschungszentrum Salzburg and the Konrad Lorenz Institute for Evolution and Cognition Research.

Philosophy of Science, 74 (January 2007) pp. 1–27. 0031-8248/2007/7401-0005$10.00

idea underlying this view is that a basic function of language is to facilitate coordinated behavior. Meaning is thus a consequence of pragmatic factors.

Lewis (1969) was intended as an answer to Quine's (1936) counterarguments against the logical positivist's claim that conventions of meaning form the basis of logical truth and logical inference. One way to express Quine's skepticism is to doubt that the truth of a statement is given by a linguistic component and a factual component (Quine 1953). Analytical statements have zero factual component. They are true regardless of how the world looks. The linguistic component of the truth of a statement is governed by conventions of meaning. But do we have a good explanation of how conventions of meaning come about?

According to Quine and others, we do not have such an explanation (see White 1950; a similar argument can already be found in Rousseau [1755]). There is no noncircular explanation of conventions of meaning. Conventions come about by agreement. To achieve agreement we always have to presuppose some kind of rudimentary language, which is itself left unexplained. One of Lewis's basic insights—which can, in fact, be traced back to Hume—is that conventions need not come about by agreement. They might be stable outcomes of repeated nonverbal interactions. Or they might be salient among their alternatives.

This paper continues Skyrms's (1996, 2000, 2004) analysis of the explanation of meaning conventions in evolutionary terms. I shall argue that an evolutionary explanation avoids the difficulties that an explanation in terms of rational choice (like Lewis's account) faces at a very fundamental level. As such, the position underlying this study is the one mentioned above, to wit, that meaning is a consequence of pragmatic factors. By this I mean that meaning emerges from the interactions of less than fully rational agents (i.e., agents who are constrained in their computing capacities and in the information they have about the structure of the game). These agents may be less deliberate than their fully rational kin in standard game theory. However, they might still be said to maximize their utility within constraints that are not assumed in standard game theory. In this sense, my analysis might be called pragmatic.

The arguments in this direction will be substantiated by reporting a number of theorems on simple signaling games and signaling games that allow probabilistic associations between states, signals, and acts. Both kinds of signaling games allow for states with no communication, states with partial communication, and states with perfect communication. My main results show that, except for one special case, the latter two types of states emerge with positive probability under selection dynamics. States with no communication are always dynamically unstable. States of partial communication can be destabilized by neutral drift. States of perfect com-

TABLE 1. A State-Act Coordination Problem with $a > 0$.

	α_1	α_2
σ_1	a	0
σ_2	0	a

munication turn out to be the only states that are robust relative to selection and drift.

I will start in Section 2 by defining simple signaling games and reviewing some of their properties. In Sections 3 and 4 I shall discuss language conventions with respect to evolution and rational choice. This discussion will be a methodological background to the study of the dynamics of simple and generalized simple signaling games in Sections 5 and 6, where the main new results can be found. Section 7 concludes by reconsidering the explanatory value of my results.

2. Simple Signaling Games. Simple signaling games are based on coordination problems between states and acts. The simplest situation of this kind consists of two states of the world, σ_1 and σ_2, and two corresponding acts, α_1 and α_2. Each act is a proper response to exactly one of the states. An individual who chooses the wrong act gets no positive payoff. This payoff structure is illustrated in Table 1. Let us call a situation like this one a "state-act coordination problem."

Suppose that there are two individuals. The sender observes the state while the receiver chooses an act. The latter cannot observe the state. The sender can send two messages, m_1 and m_2, to indicate which state has occurred. The receiver might respond to each of the signals by choosing a particular act. If the receiver chooses the right act, then both players get the same payoff a ($a > 0$). Since the sender and the receiver always get the same payoff, it is in the sender's interest to communicate which state has occurred. Likewise, the receiver has an interest to associate the signal with the right act. So they need a common understanding about the two signals in order to coordinate their actions.

The situation outlined above constitutes a simple signaling game with two players, the sender and the receiver, two states of the world, two messages, and two acts. A sender strategy specifies, for each state, what signal to send. A receiver strategy specifies which act will be chosen as a response to a message. Accordingly, there are four sender strategies and four receiver strategies. The sender might send one of the two signals if σ_1 occurs and the other one if σ_2 occurs, or she might always send the same signal regardless of which state occurs. The receiver might choose one of the two acts as a response to m_1 and the other act as a response

TABLE 2. Sender Strategies and Receiver Strategies.

s_1	m_1 if σ_1, m_2 if σ_2	r_1	α_1 if m_1, α_2 if m_2
s_2	m_2 if σ_1, m_1 if σ_2	r_2	α_2 if m_1, α_1 if m_2
s_3	m_1 if σ_1, m_1 if σ_2	r_3	α_1 if m_1, α_1 if m_2
s_4	m_2 if σ_1, m_2 if σ_2	r_4	α_2 if m_1, α_2 if m_2

to m_2, or she might ignore the message and always choose the same act (see Table 2).

The state-act coordination problem underlying our simple signaling game induces the players' payoffs. For $a = 1$, these payoffs are illustrated in Table 3. (I shall assume that $a = 1$ throughout the paper since this just amounts to a choice of scale.) Notice that we have assumed that each state occurs with probability 1/2. Each entry represents the signaler's as well as the receiver's payoff. Thus our simple signaling game is a pure coordination game (Lewis 1969).

Simple signaling games based on state-act coordination problems can easily be generalized to signaling problems involving more then two states, acts, and messages. Let $\Pi_n = \langle S, A, u^* \rangle$ be an n-state-act coordination problem if $S = \{\sigma_1, \ldots, \sigma_n\}$ is a set of n distinct states of the world, $A = \{\alpha_1, \ldots, \alpha_n\}$ is a set of n distinct acts, and u^* is a function that determines the utility of each state-act pair such that $u^*(\sigma_i, \alpha_j) = \delta_{ij}$ (where the Kronecker symbol $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise). In addition, let $M = \{m_1, \ldots, m_n\}$ be a set of n distinct messages and $\mathbb{P}$ be a probability distribution over S such that the probability of each state is positive, $\mathbb{P}(\sigma_i) > 0$ for $i = 1, \ldots, n$.

Definition 1 (simple signaling game). Let Π_n be an n-state-act coordination problem, let M be a set of n distinct messages, and let $\mathbb{P}$ be a probability distribution over S such that $\mathbb{P}(\sigma_i) > 0$ for $i = 1, \ldots, n$. A simple signaling game Σ_n based on Π_n is a triplet $\langle I, \{S_i\}_{i \in I}, \{u_i\}_{i \in I} \rangle$, where

1. $I = \{1, 2\}$ is a set of two players, the sender, 1, and the receiver, 2;
2. S_i, $i = 1, 2$, is the set of strategies generated from Π_n as follows: $S_1 = \{s_k | s_k : S \rightarrow M\}$ is the set of sender strategies and $S_2 = \{a_l | a_l : M \rightarrow A\}$ is the set of receiver strategies; and
3. the players' utility functions are the same and are generated by Π_n as follows: $u : S \times R \rightarrow \mathbb{R}$ and

$$u(s_k, r_l) = \sum_{j=1}^{n} \mathbb{P}(\sigma_j) \cdot u^*(\sigma_j, (r_l \circ s_k)(\sigma_j)),$$

where $\circ$ denotes the operation of function composition.

TABLE 3. Payoffs in a Simple Signaling Game.

	r_1	r_2	r_3	r_4
s_1	1	0	1/2	1/2
s_2	0	1	1/2	1/2
s_3	1/2	1/2	1/2	1/2
s_4	1/2	1/2	1/2	1/2

The third condition states that the payoffs to each player are generated by the underlying payoff function of the state-act coordination problem by averaging each player's payoff in each of the states σ_i according to σ_i's probability of occurrence (the second argument of u^* determines the act chosen by the receiver as a response to the signaler's message). Thus, the payoff from a particular strategy combination is the expected value of the payoffs associated with the state-act pairs that result from this strategy combination relative to the probability distribution $\mathbb{P}$.

Signaling systems are combinations of sender strategies and receiver strategies that deserve special attention (Lewis 1969). They guarantee that both players get the maximum payoff regardless of which state of the world occurs. If the players employ a signaling system, they are fully coordinated by virtue of the signals. In the example above (see Table 3), there are two signaling systems, (s_1, r_1) and (s_2, r_2).

> **Definition 2** (signaling system). Let Σ_n be a simple signaling game. Then (s_k, r_l) is a signaling system if $u(s_k, r_l) = 1$.

Equivalently, we may call (s_k, r_l) a signaling system if and only if $u^*(\sigma_j, (r_l \circ s_k)(\sigma_j)) = 1$ for all $j = 1, \ldots, n$.

Call s_i (or r_j) a part of a signaling system if and only if there is an r_k (or s_l) such that (s_i, r_k) (or (s_l, r_j)) is a signaling system. Then it is obvious that the following holds:

> **Proposition 3**. Let Σ_n be a simple signaling game. Then s_i is part of a signaling system if and only if s_i is one-to-one and r_j is part of a signaling system if and only if r_j is one-to-one.

According to Lewis (1969), Skyrms (1996), Vanderschraaf (1998), and Young (1998), a behavioral regularity is conventional if everybody has an interest to act in accordance with it and if it has an alternative (i.e., it involves some kind of arbitrariness). This intuition can be quite naturally captured in game-theoretic terms. A Nash equilibrium is a combination of strategies in which no player would gain by unilaterally deviating from her part of the equilibrium. A strict Nash equilibrium is a Nash equilibrium in which each player would do worse by unilaterally choosing a strategy different from her equilibrium strategy. In a pure coordination game with at least two strict Nash equilibria, each of the two strict Nash equilibria is a candidate for a convention. In a simple signaling game, for instance, each player does strictly better in a signaling system equilibrium and no player has an interest that the other player deviates from it since she would also be worse off in this case.[1]

In the example above (see Table 3) there are two strict Nash equilibria, (s_1, r_1) and (s_2, r_2), and four pure nonstrict Nash equilibria, (s_3, r_3), (s_3, r_4), (s_4, r_3), and (s_4, r_4). The two strict Nash equilibria are also the two signaling systems and, hence, the two possible signaling conventions of this game. This result holds with some generality (as has already been pointed out in Skyrms [1996, 83]).

> **Proposition 4**. Let Σ_n be a simple signaling game. Then (s_i, r_j) is a signaling system if and only if (s_i, r_j) is a strict Nash equilibrium.
>
> **Proof.** If (s_i, r_j) is a signaling system, then it is clear that unilateral deviation leads to a worse payoff. Conversely, if (s_i, r_j) is not a signaling system, then there always exists at least one sender strategy or one receiver strategy that yields no lower payoff (just rearrange the mappings). The details are left to the reader. ■

Thus, if conventions in signaling games are strict Nash equilibria, the only candidates for conventions in simple signaling games are signaling systems. In a population of individuals who are repeatedly playing a simple signaling game, a signaling system is a simple conventional language. The population could have adopted another signaling system that would have done essentially the same job. But whatever signaling system they have, they understand each other by virtue of a convention.

3. Stability and Emergence of Language Conventions. At this point two questions become pressing: How is a conventional language maintained in a population? And how might a conventional language be established in the first place? These two questions concern conventions in general. It is not enough to state what the candidates for conventions in a particular game are if we want to explain why one of the possible conventions is in fact a convention in a population. To do this, we have to explain how this convention has emerged and why it is stable.

These questions become particularly pressing if the candidates for conventions are entirely symmetric as they are in a simple signaling game. There is no reason to choose one of the possible signaling conventions

[1] For more on the definition of conventions in pure and impure coordination games, see Vanderschraaf (1998). Adopting an evolutionary perspective would require us to emphasize the historical process leading to a convention in its definition (see Harms 2004). This is, in some sense, achieved by the dynamical analysis below.

rather than another since any signaling system does the same job. But if there is no reason to choose one of the language conventions rather than another, how will individuals decide on one of them without communication? And, once the decision is made, why do they not switch? This scenario is one instance of what Skyrms (1996) calls the "curse of symmetry."

Lewis (1969) proposes an answer to these questions in terms of rational choice theory and in terms of salience. According to Lewis, the stability of conventions is guaranteed because the structure of the game, the payoffs, and the rationality of the players are common knowledge. A fact *F* is common knowledge among agents *A* and *A′* if *A* knows that *F*, and *A* knows that *A′* knows that *F*, and *A* knows that *A′* knows that *A* knows that *F*, and so on. The same holds if we interchange *A* and *A′*. If sender and receiver have common knowledge of the game structure and each other's rationality, they will stick to a signaling convention because there is no room for doubting that the other player is going to stick to her part of the signaling convention.

Following Schelling (1960), Lewis's answer to the question of how conventional languages are established in the first place is that one of the strict Nash equilibria is a focal point or is salient. It stands out among all equilibria in some psychologically significant respect. That is to say, it is clear to all players that this one equilibrium should be chosen. Here is Lewis's (1969, 158–159) example for a salient signaling system: Suppose that you come upon a patch of quicksand and you want to warn others who might come there after you. A salient signal would be to put a scarecrow up to the chest into the quicksand.

As Skyrms (1996) points out, both of Lewis's answers face serious problems. Concerning the first answer, Skyrms asks where all the common knowledge comes from. It seems to be unclear to what extent players must already understand each other to have this very demanding kind of knowledge. It is possible that we have to assume preexisting communication or something equivalent to it in order to explain that the players understand each other. This would put us back in the position where we need a language to explain the emergence of another language. Hence, without explaining where common knowledge comes from, Lewis's account of why conventions are stable appears to be incomplete, to say the least.

Concerning Lewis's second answer, Skyrms asks where the salient equilibrium is. Sometimes a salient equilibrium might be available, but sometimes not. The harder case is the latter one. For a fundamental investigation a solution to the less hard cases is not enough. (This criticism is a methodological one, but it does not mean that explanations in terms of salience might not be useful in other contexts.) One of the reasons why we set up simple signaling games as involving symmetric strict Nash equilibria was that no one signaling system should stand out. So, to put Skyrms's second criticism in a slightly different way, if there would be a signaling system that is salient to the players, then this should already be expressed in the structure of the game.

Allow me to add a further criticism of salience as an equilibrium selection device. As in the case of common knowledge, it seems unclear to what extent the players must already know about each other in order to view an equilibrium as salient. In terms of signaling systems this means that I must already see the signals as somewhat meaningful or representational within my community. This is, however, the fact to be explained. For instance, in the scarecrow example, one must understand a scarecrow as a representation of 'me' (or 'potentially me') and the situation as representing 'me potentially sinking'.

Cubitt and Sugden (2003) criticize Skyrms's (1996) and Vanderschraaf's (1998) reconstruction of Lewis's account of convention. First, they show that Lewis's conception of common knowledge does not coincide with the standard infinitely iterated knowledge conception (as described above). Moreover, they try to reestablish salience as an explanation of how conventions start.

Instead of assuming common knowledge, Cubitt and Sugden argue that Lewis tries to derive common knowledge from the agents' background information and their inductive standards. As they show more formally, "common knowledge in Lewis' sense is possible only when individuals have reason to believe that, in particular relevant respects, they have common background information and common inductive standards" (Cubitt and Sugden 2003, 185). Given this explanation of common knowledge, we might again ask where the reasons to believe to have common background information and common inductive standards come from, and to what extent this presupposes a mutual understanding that must itself be explained.[2]

As to the role of salience for the selection of strict Nash equilibria, Cubitt and Sugden do not address the criticisms raised above, namely, that assuming salience excludes the worst case scenario for the emergence of conventions and that the prior extent of mutual understanding for salience to be possible is unclear. Thus, it seems that Lewis's account of the emergence and stability of language conventions is deficient in two important respects. This is the point where evolutionary theory comes in.

[2] It seems to be almost inconceivable to think of a solution concept for one-shot games that does not presuppose some kind of common understanding of the situation.

4. Evolution and Conventions of Language. Skyrms (1996) invites us to approach the problem of the emergence and stability of language conventions from an evolutionary standpoint. If an adaptive process governs a population of individuals that are repeatedly confronted with a situation that is similar enough to a signaling game, will one of the signaling systems eventually become established in this population? Biological evidence suggests that the answer to this question is, at least a cautious, yes. Signal coordination is everywhere, from cells communicating via molecules and the honeybee's dance to predator alarm calls of monkeys and bird song (for comprehensive treatments of animal signals, see Snowdon [1990], Hauser [1997], and Maynard Smith and Harper [2003]).

The adoption of the evolutionary viewpoint implies that we do not follow Lewis (1969, 58) and Vanderschraaf (1998) in invoking any common knowledge assumptions to define conventions. Instead, we shall say that a population adopts a certain convention if it is a strict Nash equilibrium in a game with at least two strict Nash equilibria and if every individual in the population chooses her actions according to this strict Nash equilibrium. By adopting the evolutionary viewpoint, we also escape the criticisms raised in the previous section. We assume neither that the individuals in the population reach a convention by explicit agreement, nor that they have a preexisting language or common knowledge of the game. Indeed, they may not have much knowledge at all.

The omnipresence of signaling in nature suggests that biological and cultural evolution are likely to be responsible for these phenomena. To get beyond such informal statements, we have to give the problem a clear formulation. Studying the evolutionary dynamics of signaling games seems to be a promising starting point for a first analysis. Skyrms (1996) investigates a simple signaling game with two signals by simulating its evolutionary dynamics. In addition, he provides some analytical results in Skyrms (2000) for a simplified model of this game. In this model, a signaling system is a global attractor for a population that consists of three types, the signaling system type and two antisignaling types. On the basis of these results one might conjecture that populations will develop a signaling system under evolutionary dynamics for all simple signaling games regardless of the initial state. The results presented in the following section show that this is in general not true for the replicator equations, which are the standard dynamics in formal models of selection.

5. Replicator Dynamics and Signaling Games. If we want to study one population of individuals who can be senders or receivers, we have to consider the symmetrized, or role-conditioned, version of Σ_n (see, e.g., Cressman 2003). Denote the role-conditioned version of Σ_n by Σ_n^r. Going from Σ_n to Σ_n^r amounts to assuming that each individual is sender or receiver with probability 1/2. This is a reasonable assumption as long as we do not impose more structure on the population. If some individuals are in the role of the sender more often than others, then we must say which individuals tend to be senders and which individuals tend to be receivers. For a first analysis of the dynamics of signaling games, imposing such a structure on the population does not seem to be reasonable.

An individual's strategy consists of a sender part s_i and a receiver part r_j. If a simple signaling game is given, then the strategies of Σ_n^r are all pairs of strategies of Σ_n. (This implies that a player's sender strategy need not be consistent with her receiver strategy. She might not be able to "talk to herself.") Since Σ_n has n^n sender strategies and n^n receiver strategies, Σ_n^r has n^{2n} strategies. For instance, in the example illustrated in Table 3 there are 16 strategies. For $n = 3$, there are already 729 strategies. The payoff to a strategy (s_i, r_j) against (s_k, r_l) is given by

$$\pi((s_i, r_j), (s_k, r_l)) = \tfrac{1}{2}(u(s_i, r_l) + u(s_k, r_j)).$$

This is just the expected payoff to a strategy relative to the uniform distribution over the two roles of the game. This specifies Σ_n^r completely. There are two players. Both can be sender or receiver. Their strategies and payoffs are determined by Σ_n. Unlike Σ_n, Σ_n^r is a symmetric game. This means that the player positions are not distinguishable.

Suppose that a population consists of types of individuals in which each type corresponds to a strategy of Σ_n^r. Let $\pi(s, s')$ be the expected payoff individuals in state s get when meeting individuals of type s'. Then s is said to be evolutionarily stable if and only if $\pi(s, s) > \pi(s, s')$ for all $s' \neq s$; or if $\pi(s, s) = \pi(s, s')$ for some $s' \neq s$, then $\pi(s, s') > \pi(s', s')$ (Maynard Smith and Price 1973; Maynard Smith 1982). These conditions guarantee that a large population will not move away from state s once it has reached it. In other words, if a small proportion of individuals are not of type s in a population consisting almost entirely of s, then selection will carry the population back to a state in which only s is present.

By using a result of Selten (1980), Wärneryd (1993) shows that in signaling games, signaling systems and evolutionarily stable states coincide.

> **Proposition 5**. Let Σ_n be a simple signaling game. Then (s_i, r_j) is evolutionarily stable if and only if (s_i, r_j) is a signaling system.

Notice that Proposition 5 is an answer to the first question raised in the previous section: Why are language conventions stable? They are stable in a model like that underlying the concept of evolutionary stability because they are evolutionarily stable. In addition, they are the only evolutionarily stable states.

But why should signaling systems emerge in the first place? To answer this question, let us look at the replicator dynamics. The replicator dynamics were introduced in Taylor and Jonker (1978) to highlight the dynamical considerations underlying the evolutionary stability concept (Weibull [1995] and Hofbauer and Sigmund [1998] are comprehensive treatments of the replicator dynamics).

Let the state of the population be represented by the relative frequencies of the different types. For Σ_n^r there are $\phi(n) = n^{2n}$ different types. Thus the state of the population is represented as a vector $\mathbf{x} = (x_1, \ldots, x_{\phi(n)}) \in \Delta^{\phi(n)}$, where $\Delta^{\phi(n)}$ is the simplex of $\mathbb{R}^{\phi(n)}$:

$$\Delta^{\phi(n)} = \left\{ \mathbf{x} \in \mathbb{R}^{\phi(n)} : \sum_i x_i = 1,\ 0 \leq x_i \leq 1,\ i = 1, \ldots, \phi(n) \right\}.$$

The interior of $\Delta^{\phi(n)}$ is the part of $\Delta^{\phi(n)}$ where all types have positive frequency. It is given by

$$\operatorname{int}(\Delta^{\phi(n)}) = \left\{ \mathbf{x} \in \mathbb{R}^{\phi(n)} : \sum_i x_i = 1,\ 0 < x_i < 1,\ i = 1, \ldots, \phi(n) \right\}.$$

The boundary of $\Delta^{\phi(n)}$ is the set where at least one type has zero frequency, that is, $\operatorname{bd}(\Delta^{\phi(n)}) = \Delta^{\phi(n)} \setminus \operatorname{int}(\Delta^{\phi(n)})$. We assume that the population is effectively infinite. Thus, the actual payoff of a type matches its expected payoff. Let $\pi(x_i, \mathbf{x})$ be the average payoff of type i when the current population state is $\mathbf{x}$, and let $\pi(\mathbf{x}, \mathbf{x})$ be the average payoff of the whole population. The replicator dynamics is a system of differential equations given by

$$\dot{x}_i = x_i(\pi(x_i, \mathbf{x}) - \pi(\mathbf{x}, \mathbf{x})) \quad \text{for } i = 1, \ldots, \phi(n), \tag{1}$$

where $\dot{x}_i$ denotes the time derivative in the ith component. Hence the frequency of types with above-average payoff increases and the frequency of types with below-average payoff decreases. Equation (1) guarantees that evolution will occur in our population whenever there are fitness differences between types.

The replicator equations (1) were originally introduced to capture biological phenomena. There are, however, a number of studies that show how to make sense of the replicator equations in a cultural context (Binmore, Gale, and Samuelson 1995; Björnstedt and Weibull 1996; Schlag 1998; Harms 2004). For example, Björnstedt and Weibull show that a model in which individuals imitate others who have adopted successful strategies leads to a class of dynamics called "monotonic" by taking appropriate limits. This class of dynamics is characterized by the property that strategies yielding a higher payoff have a higher growth rate. The replicator dynamics (1) is monotonic in this sense and may be obtained by choosing specific functional forms for the functions involved in general monotonic dynamics. The system (1) may describe a learning process governed by imitation. Thus (1) seems to be a good point of departure for analyzing signaling interactions in the context of both biological and cultural evolution.

The rest points of a system of ordinary differential equations are the states at which $d\mathbf{x}/dt$ vanishes. Rest points are fixed points of the flow corresponding to the differential equations. Basically, there are three different kinds of rest points: asymptotically stable, weakly stable, and unstable rest points. An asymptotically stable rest point is characterized by attracting all nearby states. The set of all points that converge to an asymptotically stable rest point is its basin of attraction. If a rest point is weakly or Liapunov stable, then all nearby states stay nearby. (Note that every asymptotically stable rest point is weakly stable, but not vice versa.) Nearby solutions of an unstable state tend away from it (not necessarily in all directions). For details, in particular, facts concerning the relation between stability and eigenvalues, see Hirsch and Smale (1974). There is another concept we will need. A subset F of state space is called an "attractor" if all states sufficiently close to F converge to F. Thus an asymptotically stable rest point is a singleton attractor.

If $\mathbf{y}$ is an asymptotically stable state, then the system will tend back to $\mathbf{y}$ after a small perturbation. The same is true of any attractor F. If $\mathbf{y}$ is Liapunov stable, small perturbations will result in a nearby state. If $\mathbf{y}$ is unstable, then small perturbations will lead away from it. Thus asymptotically stable states are meaningful since we can expect the system to be close to one of them after a sufficiently long time.

The rest of this section is devoted to studying the stability properties of the rest points of signaling games under the replicator dynamics (1). As pointed out at the end of the last section, signaling systems will not almost surely emerge in the general case. I will show this by proving that the set of points $\mathbf{x} \in \text{int}(\Delta^{\phi(n)})$ that does not converge to a signaling system of Σ_n^r has positive Lebesgue measure. I will identify the sets of points that attract some significant portion of states as being on the boundary. Except for signaling systems, these sets are not attractors, however. Thus it will be important to be cautious when interpreting the results. After stating the relevant theorems, I will try to clarify what they mean for evolutionary explanations of signaling systems. All proofs may be found in the Appendix.

The proof of the first part of Theorem 6 relies on the fact that all rest points in $\text{int}(\Delta^{\phi(n)})$ for the evolutionary dynamics of signaling games are (linearly) unstable. Notice that Theorem 6 holds for the replicator dynamics of any symmetrized simple signaling game.

Theorem 6. Let (1) be the replicator dynamics for Σ_n^r.

1. Denote the set of points in $\text{int}(\Delta^{\phi(n)})$ that do not converge to $\text{bd}(\Delta^{\phi(n)})$ by S. Then S has Lebesgue measure zero.
2. A state $\mathbf{p}^* \in \Delta^{\phi(n)}$ is asymptotically stable if and only if $\mathbf{p}^*$ is a signaling system.

Hence, generically interior rest points will not be observed under the replicator dynamics. This situation is analogous to unstable states in statistical mechanics. For instance, the state in which all molecules in a gas are at rest is an unstable equilibrium since any slight perturbation will lead to a state that ultimately converges to mixing. If a population is at an interior rest point for the replicator dynamics of Σ_n^r, then slight perturbations will carry it to some boundary state.

For a special kind of signaling game with two signals we can in fact prove more:

Theorem 7. Let $\mathbb{P}(\sigma_1) = \mathbb{P}(\sigma_2)$. Then the set of points that do not converge to a signaling system of Σ_2^r for the replicator dynamics (1) has Lebesgue measure zero.

Thus, almost all solutions will converge to a signaling system under the assumptions of Theorem 7. The next theorem shows that the assumption $\mathbb{P}(\sigma_1) = \mathbb{P}(\sigma_2)$ is necessary for obtaining this result. The intuition for this is quite simple. If $\mathbb{P}(\sigma_1) = p > 1/2$, then the type $z = (s, r)$ in which the sender strategy s maps both states on m_1 and the receiver strategy r maps both messages on α_1 gets a payoff of p when meeting itself. There are three other types that get p when meeting each other or when meeting z. These types have r as receiver strategy and an arbitrary sender strategy. The payoff of all other types against z is at most p. Although z is unstable (it can be invaded by one of the signaling systems), a mixture of z and a small amount of the other three types with r as receiver strategy will result in a Liapunov stable state z'. The reason for this is that for any type w that is not present at z' there is at least one type v that is present at z' such that $\pi(w, v) < p$. (Details can be found in the Appendix.)

Theorem 8. Let $\mathbb{P}(\sigma_1) \neq \mathbb{P}(\sigma_2)$. Then there exist boundary faces Δ^m, $m < 16$, that contain sets of points R such that R is open in Δ^m and the set of points in Δ^{16} converging to R for the replicator dynamics (1) of Σ_2^r has positive Lebesgue measure.

This result reflects the fact that the value of p characterizes the importance of communication for a population playing Σ_2. If $p = 1/2$, then both states have equal weight. Thus coordination of both state-act pairs is equally important to obtain a sufficiently high payoff. This no longer holds for $p \neq 1/2$. In this case, coordination of the state-act pair where the state has higher probability is more important than coordination of the other state-act pair. As $p \to 1$ (or, alternatively, as $p \to 0$), communication becomes less important. If p is very close to 1, numerical simulations suggest that the measure of the set of points converging to states that are not signaling systems is nearly as large as the basins of attraction for the signaling systems.

If $n > 2$, nonconvergence to signaling systems of Σ_n does not depend on $\mathbb{P}$ as in the case of Σ_2. Moreover, for $n > 2$, not only subsets of boundary faces are attracting a significant number of initial states. Now all of the interior of a boundary face may consist of Liapunov stable states.

> **Theorem 9**. Let Σ_n^r be a simple signaling game with $n \geq 3$. Then there exist boundary faces Δ^m, $m < \phi(n)$, such that the set of points in $\Delta^{\phi(n)}$ converging to $\operatorname{int}(\Delta^m)$ for (1) has positive measure.

Let us look at one example of such a boundary face. Consider the two types $z_1 = (s_1, r)$ and $z_2 = (s_2, r)$ of Σ_3^r, where s_1 and s_2 map σ_1 and σ_2 on m_1, s_1 maps σ_3 on m_2, and s_2 maps σ_3 on m_3. r maps m_1 on α_2 and m_2 as well as m_3 on α_3. Then $\epsilon z_1 + (1 - \epsilon)z_2$ with $0 < \epsilon < 1$ consists entirely of Liapunov stable rest points, and the average payoff on this one-dimensional simplex is $\mathbb{P}(\sigma_1) + \mathbb{P}(\sigma_2) = r$. The points on this one-dimensional simplex are Liapunov stable because for any type $w \neq z_1$, z_2, $\pi(w, z_i) \leq r$ and, in addition, for all w with $\pi(w, w) > r$, $\pi(w, z_i) = r$ implies $\pi(w, z_j) < r$ for $i \neq j$ and $i, j = 1, 2$. Thus no mixture of strategies different from z_1 and z_2 can destabilize this part of the boundary as long as $0 < \epsilon < 1$.

We have to be cautious in interpreting the results stated in the previous theorems. At first sight they might appear to subvert the thesis that evolution facilitates the emergence of simple communication systems. States of communication are represented by signaling systems. These states do not almost always emerge, however, since a significant portion of initial conditions in $\Delta^{\phi(n)}$ converges to states that are not signaling systems. Taking a closer look at those suboptimal states reveals that they are not attractors. To see this, suppose that $O \subset \Delta^m$, $m < \phi(n)$, is a set of rest points open in Δ^m such that the set of interior initial conditions converging to O has positive Lebesgue measure. Then $\mathbf{p}^* \in O$ is Liapunov stable. Each trajectory starting close to $\mathbf{p}^*$ converges to it, stays close to it, or converges to a nearby rest point in O. A basic proposition in evolutionary game theory states that a point is a Nash equilibrium if it is Liapunov stable (Hofbauer and Sigmund 1998). Hence $\mathbf{p}^*$ is a Nash equilibrium. Let e and e' be pure strategies that have positive probability at $\mathbf{p}^*$. We may suppose that $\mathbf{p}^* \in \operatorname{int}(\Delta^m)$ since O is open in Δ^m. For the same reason, there exists $\epsilon \in (0, 1)$ such that

$$\pi(e, \epsilon e + (1 - \epsilon)\mathbf{p}^*) = \pi(e', \epsilon e + (1 - \epsilon)\mathbf{p}^*).$$

This implies that $\epsilon e + (1 - \epsilon)\mathbf{p}^*$ is a rest point for all $0 < \epsilon < 1$. Hence, Δ^m consists entirely of rest points. Suppose that a trajectory converges to some point $\mathbf{p}^* \in O$. Once the state of the population corresponds to $\mathbf{p}^*$, there is no selection pressure since each point in Δ^m is a rest point of the replicator dynamics. Once at $\mathbf{p}^*$, the population may visit any state in Δ^m because of neutral drift. Thus a chain of small perturbations may eventually carry the population to one of the m pure states on the boundary of Δ^m. A pure state that is not a signaling system is always dynamically unstable, however (see Wärneryd 1993). To be more specific, a pure state that is not a signaling system can always be destabilized by a signaling system. Hence, there is no neighborhood U of O such that all points in $U \cap \Delta^{\phi(n)}$ converge to some point in O. This is equivalent to saying that O is not an attractor. Thus states in O are not robust relative to drift.

In a recent paper that states a number of important complementary results to the theorems above, Pawlowitsch (2006) proposes another interpretation of the fact that the replicator dynamics can generically converge to suboptimal states. Pawlowitsch is able to show that Liapunov stable states are states of partial communication and not states with no communication at all. In particular, she shows that in the limit there can be some cases of homonymy or synonymy in the population. For simple signaling games this means that more than one state of the world may be linked to one signal and more than one signal may be linked to one action. The results on the evolutionary dynamics of signaling games may thus be seen as counterparts of inconsistencies in natural languages. Communication is not perfect in such states. But there is communication to some degree.

We might conjecture that putting mutation to the replicator equations (1) will change our results significantly. (See Hofbauer and Sigmund [1998] for more on mutation-selection dynamics.) The reason for this is that mutation often helps a population out of a local fitness maximum (which gives less payoff than a global fitness maximum like, e.g., signaling systems). Numerical simulations suggest that this is true for simple signaling games. An analytical treatment of signaling games under a mutation-selection regime is left to future research.

6. Generalized Simple Signaling Games. Signaling games may be generalized in various ways. One might, for instance, object that a more realistic model has to include probabilistic strategies. By this I mean that a sender chooses each message with some probability after the occurrence of a state and a receiver chooses an act with some probability after getting a message. Variants of this generalization have previously been studied by Oliphant and Batali (1997), Nowak and Krakauer (1999), and Komarova and Niyogi (2004).

Since not much is known about the replicator dynamics with infinite strategy spaces, I will restrict considerations to finitely many probabilistic strategies. Note that this is a crucial assumption for the subsequent analysis. It does not seem to be crucial for the result, however. A natural extension of the construction of the strategy space for generalized simple signaling games that allows a continuum of strategies would be desirable.

The sender strategies are now given by a probability distribution over the set of messages M for each state in S. Similarly, a receiver strategy is given by a probability distribution over the set of acts A for each message in M. If individuals are allowed such probabilistic strategies, then a strategy of a generalized simple signaling game is a mixed strategy of Σ_n^r; that is, it corresponds to a point $\mathbf{x} \in \Delta^{\phi(n)}$. Suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \Delta^{\phi(n)}$ are the mixed types under consideration. The payoffs of $\mathbf{x}_i$ against $\mathbf{x}_j$ are given by $b_{ij} = \mathbf{x}_i \cdot A\mathbf{x}_j$, where A is the payoff matrix defined for Σ_n^r. Since A is symmetric, $b_{ij} = b_{ji}$. Hence, the payoff matrix for the generalized simple signaling game B is symmetric. This, in turn, implies that the results presented in the previous section basically carry over to generalized simple signaling games. (The symmetry of the payoff matrix B allows the same analysis of the dynamics of generalized simple signaling games as the one given in the Appendix for simple signaling games.) Suppose that Σ_2 has two equiprobable states. If the two signaling systems of Σ_2^r are present among the probabilistic strategies $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \Delta^{16}$ and if m is finite, then the set of initial states that do not converge to one of the signaling systems has Lebesgue measure zero. Suppose that Σ_2 has two states with unequal probabilities of occurrence. Then a significant number of states do not converge to one of the signaling systems. If $n \geq 3$, then this result is independent of the probability distribution over the set of states. For all generalized simple signaling games, almost every initial state converges to the boundary and signaling systems are the only attractors.

7. Conclusion. To what extent have studies like this one or Skyrms (1996) been able to answer Quine's skeptical doubts concerning conventional meaning? There are several points to be noted.

In the first place, I chose to explain the spontaneous emergence of meaning with the help of evolutionary theory. To do this in a specific way, I worked with a particular model, the replicator dynamics. For signaling games, the explanatory value of signaling system equilibria depends on the stability properties of the corresponding rest points. I have shown that signaling systems are the most robust states for the replicator dynamics. All the other states are not robust under selection or relative to neutral drift. This should not cover up the fact that there exist sets of states that attract a significant part of state space. These sets consist of states in which there is some, but not perfect, communication among individuals. Thus it may be quite hard for a population to get to a state of (almost) perfect communication under selection dynamics. There is reason to suppose, however, that states of partial communication will be unstable once mutation is explicitly introduced into the dynamics. States of partial communication can also be seen as corresponding to persistent imperfections of natural languages.

One way to improve the significance of these results is to study whether they continue to hold in more general models. I have taken one step in this direction by showing that the stability results still hold for a much bigger strategy space with randomized strategies. One might object that for addressing the question of how meaning emerges, an even bigger strategy space has to be considered. More specifically, strategies in which agents do not use any signal and never react to receiving a signal should be included. This might be considered as a more realistic starting point for an initial population state. In one sense, simple and generalized simple signaling games include such a state. As long as signals are costless, being silent can itself be regarded as a signal and thus be included in the set of signals. Similarly, the act of just carrying on doing what one is doing at the time of receiving a message can be a member of the set of acts. After this particular signal and this particular act are specified, the strategies of never using a signal and never reacting to signals are part of the strategy space. But such a state is not stable in our models. This result might not hold in more complex models, however. Signals may be costly, for instance, because of the capacities an individual must possess in order to be able to produce signals and to learn how to produce them. In this case, sending no signal at all might be advantageous if many other individuals never react to signals. The outcome will depend on the cost of being able to send signals and on the benefits from communication.

Investigating the robustness of results by studying different models is another important issue. Skyrms (2000) employs considerations of structural stability, where small perturbations in the differential equations are studied, and by looking at a bigger class of dynamics, qualitatively adaptive dynamics, which includes the replicator dynamics. His results suggest that we might be able to extend our general analysis in this direction. Furthermore, there exist some simulation studies on signaling games played in a spatial environment (Grim et al. 2001; Zollman 2005). The general result in this direction is that signaling systems are likely to emerge.

Another point should be emphasized. Signaling games can solve only part of the problem posed by Quine. After all, not only did Quine (1936, 1953) question the possibility of a noncircular account of the explanation of meaning. More generally, he questioned a conventionalist account of logic in which logical truth and logical inference are based on conventional meaning. These considerations lead to more sophisticated signaling games, some of which where proposed in Skyrms (2004). A satisfactory conventionalist approach to language cannot get around taking steps along these lines.

Finally, one might object that the meaning of signals in simple signaling games remains unclear. If an animal gives an alarm call, does it mean that a predator is present? Or does it mean that the proper response to the presence of the predator should be performed? That is to say, a Quinean skeptic might still question that we have explained conventional meaning if our explanation is not based on a model of signaling in which the meaning of signals is sufficiently close to what we commonly understand by meaning in human languages. An answer to this criticism is proposed in Huttegger (2007).

Appendix: Proofs

Let me first introduce some definitions and some notational conventions. A gradient system on an open set $U \subset \mathbb{R}^n$ is a system of differential equations of the form

$$\frac{d\mathbf{x}}{dt} = \nabla V(\mathbf{x}).$$

$V : U \to \mathbb{R}$ is assumed to be a function with continuous second-order partial derivatives and is called the “potential” of the gradient system. If the gradient ∇V is defined with respect to the standard inner product for $\mathbb{R}^n$, then

$$\nabla V = \left(\frac{\partial V}{\partial x_1}, \ \ldots, \ \frac{\partial V}{\partial x_n}\right).$$

If $\mathbb{R}^n$ is equipped with an arbitrary inner product, the gradient ∇V can straightforwardly be defined by considering the dual vector space of $\mathbb{R}^n$ (i.e., the space of all linear maps from $\mathbb{R}^n$ to $\mathbb{R}$; see Hirsch and Smale [1974] for details). The gradient systems for the replicator dynamics are “Shashshahani gradients.” This means that they are defined with respect to the following inner product:

$$\langle \xi, \eta \rangle_{\mathbf{x}} = \sum_{i=1}^{n} \frac{1}{x_i} \xi_i \eta_i.$$

Here, $\mathbf{x} \in \Delta^n$ and $\xi, \eta \in \mathbb{R}^n_0$, where $\mathbb{R}^n_0 = \{\xi \in \mathbb{R}^n : \sum_i \xi_i = 0\}$ is the tangent space for every point in Δ^n. A symmetric game Γ with payoff matrix A is a "partnership game" if $A = A^T$ (where A^T denotes the transpose of A). If $s_k, \ldots, s_{k+m}$ are strategies of a game Γ, then span$(s_k, \ldots, s_{k+m})$ denotes the set of all convex combinations of those strategies.

Before I prove the theorems stated in the main text, I will first prove two lemmata that will be used frequently below.

Lemma 10. Let Γ be a partnership game with $n \times n$ payoff matrix A. Then:

1. $\mathbf{x} \in S^n$ is evolutionarily stable if and only if $\mathbf{x} \in S^n$ is asymptotically stable under the replicator dynamics (1) generated by A.
2. The replicator dynamics (1) for Γ is a Shashshahani gradient system with potential function $V(\mathbf{x}) = (1/2)\mathbf{x} \cdot A\mathbf{x}$.

Proof. See Hofbauer and Sigmund (1998). ■

The second part of Lemma 10 implies that all solutions converge to a rest point (Akin and Hofbauer 1982). Moreover, it implies that there are no circling solutions since the average payoff V is strictly increasing along all nonstationary solutions.

Lemma 11. Let Σ_n be a simple signaling game. Then the role-conditioned game based on Σ_n where each player finds herself in the sender position as well as in the receiver position with probability 1/2 is a partnership game.

Proof. We have to show that the payoff matrix of the role-conditioned game based on Σ_n is symmetric. Consider two individuals of type $z = (s_i, r_j)$ and $z' = (s_l, r_k)$. Since we have supposed that any individual finds herself in each of the positions with probability 1/2,

$$\pi(z, z') = \tfrac{1}{2}[u(s_i, r_j) + u(s_l, r_k)].$$

It is easy to see that the payoff to z' must be the same. ■

Proof of Theorem 6

The second part of Theorem 6 follows directly from Proposition 5, the first part of Lemma 10, and Lemma 11. The first part of Theorem 6 follows from the three lemmata stated below together with the second part of Lemma 10.

Lemma 12. Let Σ_n^r be a simple signaling game. If $\mathbf{p}^* \in \text{int}(\Delta^{\phi(n)})$ is a rest point of (1), then $\mathbf{p}^*$ is linearly unstable.

Proof. Let $\mathbf{p}^*$ be an interior rest point of (1). Let $p_i = p_i^* + \xi_i$, where $\xi = (\xi_1, \ldots, \xi_{\phi(n)}) \in \mathbb{R}_0^{\phi(n)}$. Then the linearized vector field around $\mathbf{p}^*$ is given by

$$\dot{\xi}_i = \sum_j L_{ij}\xi_j$$

(see, e.g., Cressman 2003). $L_{ij} = p_i^*(a_{ij} - \mathbf{p}^* \cdot Ae_j)$ is the jth partial derivative of the ith equation of (1), and A is the symmetric payoff matrix of Σ_n^r with entries a_{ij}. The components L_{ij} constitute the Jacobian matrix L evaluated at $\mathbf{p}^*$. Since Σ_n^r is a partnership game by Lemma 11, L is a self-adjoined linear operator relative to the Shashshahani inner product: $\langle \xi, L\eta\rangle_{\mathbf{p}^*} = \langle L\xi, \eta\rangle_{\mathbf{p}^*}$ (see Hofbauer and Sigmund 1998, 259). By a basic result in linear algebra, a self-adjoined linear operator is similar to a symmetric matrix in an orthonormal basis. This implies that L has at least one positive eigenvalue if and only if L is not negative semidefinite. L is not negative semidefinite if there exists some $\xi \in \mathbb{R}_0^{\phi(n)}$ such that $\langle \xi, L\xi\rangle_{\mathbf{p}^*} > 0$. Set $\xi = \mathbf{p} - \mathbf{p}^*$, where

$$\mathbf{p} = (1-\epsilon)\mathbf{p}^* + \epsilon\mathbf{s}, \quad 0 < \epsilon < 1.$$

Then

$$\langle \xi, L\xi\rangle_{\mathbf{p}^*} = \sum_{i=1}^{\phi(n)} \xi_i a_{ij}\xi_j - \sum_{i=1}^{\phi(n)} \xi_i \sum_{j=1}^{\phi(n)} \mathbf{p}^* \cdot Ae_j\xi_j$$
$$= \sum_{i=1}^{\phi(n)} \xi_i a_{ij}\xi_j = (\mathbf{p} - \mathbf{p}^*) \cdot A(\mathbf{p} - \mathbf{p}^*) = \pi(\mathbf{p}, \mathbf{p}) - \pi(\mathbf{p}^*, \mathbf{p}^*).$$

Since $\mathbf{p}^*$ is an interior equilibrium,

$$\pi(e_i, \mathbf{p}^*) = \pi(e_j, \mathbf{p}^*) = \pi(\mathbf{p}^*, \mathbf{p}^*)$$

for all pure types e_i, e_j. At $\mathbf{p}$ the payoff to $\mathbf{s}$ is

$$\pi(\mathbf{s}, \mathbf{p}) = \epsilon\pi(\mathbf{s}, \mathbf{s}) + (1-\epsilon)\pi(\mathbf{s}, \mathbf{p}^*)$$

and the average payoff is

$$\pi(\mathbf{p}, \mathbf{p}) = \epsilon\pi(\mathbf{p}, \mathbf{s}) + (1-\epsilon)\pi(\mathbf{p}, \mathbf{p}^*) > \pi(\mathbf{p}^*, \mathbf{p}^*)$$

since $\pi(\mathbf{s}, \mathbf{s}) > \pi(\mathbf{p}^*, \mathbf{p}^*)$. Thus L is not negative semidefinite. ■

A set of points S in $\mathbb{R}^n$ is called path-connected if for all points $\mathbf{x}$ and $\mathbf{y}$ in S there exists a continuous path ϕ connecting $\mathbf{x}$ and $\mathbf{y}$ and lying entirely in S.

Lemma 13. Let Σ_n^r be a symmetrized simple signaling game and let

U be a connected set of rest points in int($\Delta^{\phi(n)}$). Then there exists at most one such set U.

Proof. If $\mathbf{p}^*$, $\mathbf{q}^* \in$ int($\Delta^{\phi(n)}$) are rest points, then $\pi(e, \mathbf{p}^*) = \pi(e', \mathbf{p}^*)$ and $\pi(e, \mathbf{q}^*) = \pi(e', \mathbf{q}^*)$ for all pure strategies e, e' (since all strategies are present at an interior rest point). Then $\epsilon\mathbf{p}^* + (1 - \epsilon)\mathbf{q}^*$ is also a rest point for $0 \leq \epsilon \leq 1$ since $\pi(e, \epsilon\mathbf{p}^* + (1 - \epsilon)\mathbf{q}^*) = \pi(e', \epsilon\mathbf{p}^* + (1 - \epsilon)\mathbf{q})$ by multilinearity of π. This shows that U is a linear manifold. ■

Lemma 13 shows that the set of interior rest points is indeed connected since the line connecting two arbitrary interior rest points consists entirely of interior rest points. The proof of the next lemma is based on results from center manifold theory. Center manifold theory asserts that, at each rest point, there exist invariant manifolds tangent to the (generalized) eigenspaces spanned by eigenvectors corresponding to eigenvalues with positive, negative, and zero real part. The center-stable manifold at some rest point is the invariant manifold tangent to the union of the eigenspaces given by eigenvalues with nonpositive real part. The center-stable manifold thus contains all solutions converging to the rest point or staying sufficiently close to it.

Lemma 14. Let U be a path-connected set of interior rest points for the replicator dynamics of Σ_n^r and let S be the set of points that converge to U. Then $U \cup S$ has Lebesgue measure zero.

Proof. From Lemma 12 we know that every $\mathbf{p}^* \in U$ is linearly unstable. This together with the center-stable manifold theorem (see Kelley 1967) implies that $U' \cup S'$ is contained in the center-stable manifold $M_{\mathbf{p}^*}$ at $\mathbf{p}^*$, where U' is a set of rest points sufficiently close to $\mathbf{p}^*$ and S' is the set of points that converge to a rest point close to $\mathbf{p}^*$ and are themselves sufficiently close to $\mathbf{p}^*$ (from the second part of Lemma 10 we know that every solution converges to some rest point). Locally a center-stable manifold $M_{\mathbf{p}^*}$ exists for each interior rest point $\mathbf{p}^*$. Since each $\mathbf{p}^*$ is linearly unstable, the center-stable manifold theorem implies that $M_{\mathbf{p}^*}$ has Lebesgue measure zero. $U \cup S$ is the union of all local center-stable manifolds. Because it is a subset of $\mathbb{R}^{\phi(n)}$, Lindelöf's theorem implies that any open covering of $U \cup S$ has a countable subcovering. Since each center-stable manifold has Lebesgue measure zero and any countable union of measure zero sets has again measure zero, this implies that $U \cup S$ has Lebesgue measure zero. ■

Proof of Theorem 7

From Theorem 6 we know that solutions starting from almost every initial condition converge to the boundary and that the two signaling systems of Σ_2 are the only asymptotically stable states of the replicator dynamics for Σ_2^r. To analyze the stability properties of boundary rest points, we look at the following linear manifolds:

- $\Delta_1 = \text{span}(\{z : z = (s_i, r_j)$ and r_j is not one-to-one$\})$.
- $\Delta_2 = \text{span}(\{z : z = (s_i, r_k)$ and s_i is not one-to-one$\})$.
- Let Δ_3 be the union of all boundary faces spanned by strategies z where at least one sender part is one-to-one, at least one receiver part is one-to-one, and at least one type is not present.

Since $\mathbb{P}(\sigma_1) = \mathbb{P}(\sigma_2)$, the average payoff is 1/2 on Δ_1 and Δ_2. Hence Δ_1 and Δ_2 consist entirely of rest points. Suppose $\mathbf{p}^* \in \Delta_1$. Let $z = (s, r)$ and $z' = (s', r')$ be the signaling systems of Σ_2^r. Let $S_1 = \{(s_i, r_j) :\ s_i$ and r_j are not one-to-one$\}$. Similarly, let $S_2 = \{(s_i, r_j) : s_i = s$ and r_j is not one-to-one$\}$ and let $S_3 = \{(s_i, r_j) : s_i = s'$ and r_j is not one-to-one$\}$. Then $\Delta_1 = \text{span}(S_1 \cup S_2 \cup S_3)$. If $z_i \in S_1$, then $\pi(z, z_i) = 1/2$. If $z_i \in S_2$, then $\pi(z, z_i) = 3/4$. And if $z_i \in S_3$, then $\pi(z, z_i) = 1/4$. Define $\alpha = \sum_{i\in S_1} p_i^*$, $\beta = \sum_{i\in S_2} p_i^*$, and $\gamma = \sum_{i\in S_3} p_i^*$. Let L denote the Jacobian evaluated at $\mathbf{p}^*$. The entries of the Jacobian for all types $e_i \notin \text{supp}(\mathbf{p}^*)$ are of the form $\delta_{ij}(\pi(e_i, \mathbf{p}^*) - \pi(\mathbf{p}^*, \mathbf{p}^*))$ (see, e.g., Cressman 2003, 51). This implies that the eigenvalue corresponding to z is given by

$$\pi(z, \mathbf{p}^*) - \pi(\mathbf{p}^*, \mathbf{p}^*).$$

If $\pi(z, \mathbf{p}^*) - 1/2 > 0$, then $\mathbf{p}^*$ is linearly unstable. If $\pi(s, \mathbf{p}^*) - 1/2 = 0$, then a simple computation of $\pi(\mathbf{x}, \mathbf{x})$ when perturbed toward z shows that $\mathbf{p}^*$ is second-order unstable. Suppose that $\pi(s, \mathbf{p}^*) - 1/2 < 0$. Then $\gamma > \beta$. This implies, however, that

$$\pi(z', \mathbf{p}^*) = \tfrac{1}{2}\alpha + \tfrac{1}{4}\beta + \tfrac{3}{4}\gamma > \tfrac{1}{2}.$$

Thus, $\mathbf{p}^*$ is again linearly unstable. A similar argument can be given for the case $\mathbf{p}^* \in \Delta_2$.

Suppose now that $\mathbf{p}^* \in \Delta_3$ is a rest point. Observe first that if at least one signaling system is present on the boundary face under consideration, then the same argument as in the proof of Lemma 12 shows that $\mathbf{p}^*$ is linearly unstable. Thus we may assume that no signaling system is in the support of $\mathbf{p}^*$. Let S_1, S_2, and S_3 be the same as above. Define in addition $S_4 = \{(s_i, r_j) :\ s_i$ is not one-to-one and $r_j = r\}$ and $S_5 = \{(s_i, r_j) :\ s_i$ is not one-to-one and $r_j = r'\}$. Let α, β, γ, δ, and ϵ be the corresponding sums of frequencies in S_i, $i = 1, \ldots, 5$, at $\mathbf{p}^*$. Then

$$\pi(z, \mathbf{p}^*) = \tfrac{1}{2}\alpha + \tfrac{3}{4}(\beta + \delta) + \tfrac{1}{4}(\gamma + \epsilon)$$

and

$$\pi(z', \mathbf{p}^*) = \tfrac{1}{2}\alpha + \tfrac{3}{4}(\gamma + \epsilon) + \tfrac{1}{4}(\beta + \delta).$$

Since $\mathbf{p}^* \in \Delta_3$, we may suppose without loss of generality that $\beta > 0$. Let $z_i \in S_2$. Then

$$\pi(z_i, \mathbf{p}^*) = \tfrac{1}{2}(\alpha + \beta + \gamma) + \tfrac{3}{4}\delta + \tfrac{1}{4}\epsilon.$$

If $\beta > \gamma$, then $\pi(z, \mathbf{p}^*) > \pi(z_i, \mathbf{p}^*) = \pi(\mathbf{p}^*, \mathbf{p}^*)$. If $\gamma > \beta$, then there exists a $z_j \in S_3$ such that $\pi(z', \mathbf{p}^*) > \pi(z_j, \mathbf{p}^*) = \pi(\mathbf{p}^*, \mathbf{p}^*)$. In both cases the Jacobian at $\mathbf{p}^*$ has a positive eigenvalue. If $\beta = \gamma$, then $\mathbf{p}^*$ is second-order unstable, which can again be seen by computing the average payoff when $\mathbf{p}^*$ is slightly perturbed in the direction of one signaling system.

This shows that almost all trajectories will converge to one of the signaling systems. To complete the proof of Theorem 7, observe that the vector field (1) is invariant under permutation of the signaling systems since the differential equations remain the same. Thus, the basins of attraction must be of equal size. ■

Proof of Theorem 8

To prove the theorem, let us first suppose a boundary face consisting of rest points for the replicator dynamics. Consider the following sender strategies: s_1 maps both states on m_1. $s_2 = s$ and $s_3 = s'$, where s and s' are the two one-to-one sender strategies. And s_4 maps both states on m_2. Define $z_1 = (s_1, r_1)$, $z_2 = (s_2, r_1)$, $z_3 = (s_3, r_1)$, and $z_4 = (s_4, r_1)$, where r_1 is the receiver strategy that maps both messages on α_1. Let $\Delta = \text{span}(z_1, z_2, z_3, z_4)$ and set $\mathbf{p}^* = \alpha z_1 + \beta z_2 + \gamma z_3 + \delta z_4$, where $\delta = 1 - \alpha - \beta - \gamma$. Suppose without loss of generality that $\mathbb{P}(\sigma_1) = p > 1/2$. Then the average payoff on Δ is p. Hence all points $\mathbf{p} \in \Delta$ are rest points.

Let $\mathbf{p}^* \in \text{int}(\Delta)$. As in the proof of Theorem 7, the eigenvalues concerning strategies $z_i \notin \text{supp}(\mathbf{p}^*)$ are given by $\pi(z_i, \mathbf{p}^*) - p$. I will show that, for some $\mathbf{p}^* \in \text{int}(\Delta)$, all these eigenvalues are negative. Consider strategies $z_i = (s_k, r_l) \notin \text{supp}(\mathbf{p}^*)$ such that r_l is many-to-one. These strategies earn a payoff of $1/2(p + q) = 1/2$ against $\mathbf{p}^*$. Since $p > 1/2$, all eigenvalues with respect to these types are negative.

Consider the following sets: $S_1 = \{(s_k, r_l) : r_l = r\}$ and $S_2 = \{(s_k, r_l) : r_l = r'\}$, where $z = (s, r)$ and $z' = (s', r')$ are again the two signaling system strategies. Then for $z_i \in S_i$,

$$\pi(z_1, \mathbf{p}^*) = \tfrac{1}{2}(1 + \alpha(2p - 1) + \beta p + \gamma(p - 1))$$

and

$$\pi(z_2, \mathbf{p}^*) = \tfrac{1}{2}(2p + \alpha(1 - 2p) - \beta p + \gamma(1 - p)).$$

If both of these expressions are negative for certain values of α, β, and γ, then eigenvalues relative to strategies in S_1 and S_2 are also negative. Solving for the corresponding inequalities shows that the set of values for α, β, and γ is nonempty for some values of $1/2 < p < 1$. The location of this set will depend on p. Suppose, for example, that

$$\frac{\alpha + \gamma - 1}{2\alpha + \beta + \gamma - 2} < p < 1.$$

If $\beta - \gamma \geq 0$, then the fraction on the left will be less than or equal to 1/2. Hence this condition on p will hold. Thus points in the set given by $0 < \alpha < 1$, $0 < \beta < 1 - \alpha$, and $0 \leq \gamma \leq \beta$ have negative eigenvalues relative to all types in S_1 and S_2, as the reader may easily verify. Thus we have only zero eigenvalues and negative eigenvalues at a suitably chosen point $\mathbf{p}^*$. The zero eigenvalues correspond to the center manifold at $\mathbf{p}^*$. The center manifold coincides with Δ. This implies that $\mathbf{p}^*$ is Liapunov stable since nearby solutions either stay nearby (in the center manifold) or approach nearby points. From this we may conclude that a set of positive measure converges to an open set on the boundary containing $\mathbf{p}^*$. ■

Proof of Theorem 9

Let us first suppose that we have a signaling game Σ_3 with $\mathbb{P}(\sigma_1) = p$, $\mathbb{P}(\sigma_2) = q$, $\mathbb{P}(\sigma_3) = 1 - p - q$, and $p = q = 1/3$. Consider the following sender and receiver strategies:

1. $\sigma_1 \mapsto m_1$ and $\sigma_2 \mapsto m_1$; $\sigma_3 \mapsto m_2$ (s_1) or $\sigma_3 \mapsto m_3$ (s_2);
2. $m_1 \mapsto \alpha_1$ (r_1) or $m_1 \mapsto \alpha_2$ (r_2); $m_2 \mapsto \alpha_3$ and $m_3 \mapsto \alpha_3$.

(1) defines two sender strategies, s_1 and s_2, while (2) defines two receiver strategies, r_1 and r_2. Let $z_1 = (s_1, r_1)$, $z_2 = (s_1, r_2)$, $z_3 = (s_2, r_1)$, and $z_4 = (s_2, r_2)$. Let $\Delta = \text{span}(z_1, z_2, z_3, z_4)$. A straightforward computation yields that the average payoff on Δ is 2/3. Hence Δ consists entirely of rest points. The corners z_i can obviously be destabilized by signaling systems (this follows again from a result in Wärneryd [1993]). Thus we suppose that $\mathbf{p}^* \in \text{int}(\Delta)$. I will show that $\mathbf{p}^*$ is quasi-strict. A strategy $\mathbf{q}$

is quasi-strict if $\mathrm{BR}(\mathbf{q}) \in \Delta(\mathrm{supp}(\mathbf{q}))$ (a best response to $\mathbf{q}$ puts positive weight on a pure strategy if and only if $\mathbf{q}$ does). The quasi strictness of $\mathbf{p}^*$ implies that all real parts of the eigenvalues corresponding to states not on Δ are negative (this follows again from the fact that the entries of the Jacobian at $\mathbf{p}^*$ are of the form $L_{ij} = \delta_{ij}(u(z_i, \mathbf{p}^*) - u(\mathbf{p}^*, \mathbf{p}^*))$ if $z_i \notin$ supp(Δ)). This together with the fact that the zero eigenvalues correspond to a manifold of rest points, Δ, implies that $\mathbf{p}^*$ is Liapunov stable. Hence, the open set int(Δ) attracts a set of positive measure from the interior of Δ^{729}.

First observe that for every sender strategy s_i with $i \neq 1, 2$, $u(s_i, r_j) \leq 2/3$ for $j = 1, 2$ since coordination with each r_j is possible only for two state-act pairs and all three state-act pairs are equiprobable. For every r_j with $j \neq 1, 2$, $u(s_i, r_j) \leq 2/3$ for $i = 1, 2$ since both sender strategies map σ_1 and σ_2 on m_1. Moreover, if $u(s_i, r_j) = 2/3$, then $u(s_k, r_j) < 2/3$ for $i \neq k$. To see this, suppose without loss of generality that $u(s_1, r_j) = 2/3$. Then r_j must map m_2 on α_3 and m_1 on α_1 or on α_2. Then, for $r_j \neq r_k$ for $k = 1, 2$, it is necessary that r_j does not map m_3 on α_3. This implies that $u(s_2, r_j) < 2/3$. Now let $z = (s_i, r_j)$ be an arbitrary strategy of Σ_n^r with $i, j \neq 1, 2$. Then

$$\pi(z, z_i) = \tfrac{1}{2}(u(s_i, r_k) + u(s_l, r_j)) \leq \tfrac{2}{3}.$$

Our arguments show moreover that there exists at least one $z_r \in$ supp($\mathbf{p}^*$) such that $u(z, z_r) < 2/3$. Since $\mathbf{p}^* \in$ int(Δ), it follows that $\pi(z, \mathbf{p}^*) < 2/3$ for all $z \notin$ supp($\mathbf{p}^*$).

Suppose $p > q$. Then similar reasoning applies to $\mathbf{p}^* \in$ int(span(z_1, z_2)). $p > q$ now implies that $u(s_i, r_2) < u(s_i, r_1)$ for $i = 1, 2$. Now r_1 is the unique receiver strategy that is optimal for both s_1 and s_2. This shows that the conclusion of Theorem 9 also holds in the case in which the three states are not equiprobable.

Suppose now that $n \geq 4$. Then similar reasoning applies to the strategies based on the following sender and receiver strategies:

1. $\sigma_1 \mapsto m_1$ and $\sigma_2 \mapsto m_1$; $\sigma_3 \mapsto m_2$ (s_1) or $\sigma_3 \mapsto m_3$ (s_2); $\sigma_4 \mapsto m_4, \ldots, \sigma_n \mapsto m_n$.
2. $m_1 \mapsto \alpha_1$ (r_1) or $m_1 \mapsto \alpha_2$ (r_2); $m_2 \mapsto \alpha_3$ and $m_3 \mapsto \alpha_3$; $m_4 \mapsto \alpha_4, \ldots, m_n \mapsto \alpha_n$.

The average payoff on Δ' defined by the four strategies based on (1) and (2) will again be constant. $u(z, \mathbf{p}^*) < u(\mathbf{p}^*, \mathbf{p}^*)$ for $\mathbf{p}^* \in \Delta'$ and $z \notin$ supp($\mathbf{p}^*$) by similar arguments as in the case of $n = 3$. ■

REFERENCES

Akin, Ethan, and Josef Hofbauer (1982), "Recurrence of the Unfit," *Mathematical Biosciences* 61: 51–62.

Binmore, Ken, John Gale, and Larry Samuelson (1995), "Learning to Be Imperfect: The Ultimatum Game," *Games and Economic Behavior* 8: 56–90.

Björnstedt, Johan, and Jörgen Weibull (1996), "Nash Equilibrium and Evolution by Imitation," in Kenneth J. Arrow, Enrico Colombatto, Mark Perlman, and Christian Schmidt (eds.), *The Rational Foundations of Economic Behaviour*. New York: Macmillan, 135–171.

Cressman, Ross (2003), *Evolutionary Dynamics and Extensive Form Games*. Cambridge, MA: MIT Press.

Cubitt, Robin P., and Robert Sugden (2003), "Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory," *Economics and Philosophy* 19: 175–210.

Grim, Patrick, Trina Kokalis, Ali Tafti, Nicholas Kilb, and Paul St. Denis (2001), *Making Meaning Happen*. Technical Report 01-02. Stony Brook: SUNY, Stony Brook, Group for Logic and Formal Semantics.

Harms, William F. (2004), *Information and Meaning in Evolutionary Processes*. Cambridge: Cambridge University Press.

Hauser, Marc D. (1997), *The Evolution of Communication*. Cambridge, MA: MIT Press.

Hirsch, Morris W., and Stephen Smale (1974), *Differential Equations, Dynamical Systems, and Linear Algebra*. Orlando, FL: Academic Press.

Hofbauer, Josef, and Karl Sigmund (1998), *Evolutionary Games and Population Dynamics*. Cambridge: Cambridge University Press.

Hume, David (1739), *A Treatise of Human Nature*. London: John Noon.

Huttegger, Simon (2007), "Evolutionary Explanations of Indicatives and Imperatives," *Erkenntnis* 66: 409–436.

Kelley, Al (1967), "The Stable, Center-Stable, Center, Center-Unstable, Unstable Manifolds," *Journal of Differential Equations* 3: 546–570.

Komarova, Natalia, and Partha Niyogi (2004), "Optimizing the Mutual Intelligibility of Linguistic Agents in a Shared World," *Artificial Intelligence* 154: 1–42.

Lewis, David (1969), *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

Maynard Smith, John (1982), *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Maynard Smith, John, and David Harper (2003), *Animal Signals*. Oxford: Oxford University Press.

Maynard Smith, John, and George Price (1973), "The Logic of Animal Conflict," *Nature* 146: 15–18.

Nowak, Martin A., and David C. Krakauer (1999), "The Evolution of Language," *Proceedings of the National Academy of Sciences* 96: 8028–8033.

Oliphant, Michael, and John Batali (1997), "Learning and the Emergence of Coordinated Communication," *Center for Research on Language Newsletter* 11. http://www.isrl.uiuc.edu/amag/langev/paper/oliphant97learningAnd.html.

Pawlowitsch, Christina (2006), "Why Evolution Does Not Always Lead to an Optimal Signaling System." Working paper, University of Vienna.

Quine, Willard V. O. (1936), "Truth by Convention," in Otis H. Lee (ed.), *Philosophical Essays for A. N. Whitehead*. New York: Longmans, 90–124.

——— (1953), "Two Dogmas of Empiricism," in *From a Logical Point of View*. Cambridge, MA: Harvard University Press, 20–46.

Rousseau, Jean-Jacques ([1755] 1997), *Discourse on the Origin and the Foundations of Inequality among Men*. Reprinted in V. Gourevitch (ed.), *The Discourses and Other Early Political Writings*. Cambridge: Cambridge University Press, 111–222.

Schelling, Thomas (1960), *The Strategy of Conflict*. Oxford: Oxford University Press.

Schlag, Karl (1998), "Why Imitate, and If So How? A Bounded-Rational Approach to the Multi-armed Bandit," *Journal of Economic Theory* 78: 130–156.

Selten, Reinhard (1980), "A Note on Evolutionary Stable Strategies in Asymmetrical Animal Conflicts," *Journal of Theoretical Biology* 84: 93–101.

Skyrms, Brian (1996), *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

——— (2000), "Stability and Explanatory Significance of Some Simple Evolutionary Models," *Philosophy of Science* 67: 94–113.

——— (2004), *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Smith, Adam ([1761] 2000), *Considerations Concerning the First Formation of Languages*. Reprinted in *The Theory of Moral Sentiments*. Amherst, NY: Prometheus, 505–538.

Snowdon, Charles T. (1990), "Language Capacities of Nonhuman Animals," *Yearbook of Physical Anthropology* 33: 215–243.

Taylor, Peter D., and Leo Jonker (1978), "Evolutionarily Stable Strategies and Game Dynamics," *Mathematical Biosciences* 40: 145–156.

Vanderschraaf, Peter (1998), "Knowledge, Equilibrium and Convention," *Erkenntnis* 49: 337–369.

Wärneryd, Karl (1993), "Cheap Talk, Coordination, and Evolutionary Stability," *Games and Economic Behavior* 5: 532–546.

Weibull, Jörgen (1995), *Evolutionary Game Theory*. Cambridge, MA.: MIT Press.

White, Morton (1950), "The Analytic and the Synthetic: An Untenable Dualism," in Sidney Hook (ed.), *John Dewey: Philosopher of Science and Freedom*. New York: Dial, 316–330.

Young, H. Peyton (1998), *Individual Strategy and Social Structure: An Evolutionary Theory of Institutions*. Princeton, NJ: Princeton University Press.

Zollman, Kevin J. S. (2005), "Talking to Neighbors: The Evolution of Regional Meaning," *Philosophy of Science* 72: 69–85.