1

An Introduction to Game Theory for Linguists

Anton Benz, Gerhard Jäger and Robert van Rooij

1 Classical game theory

In a very general sense we can say that we play a game together with other people whenever we have to decide between several actions such that the decision depends on the choice of actions by others and on our preferences over the ultimate results. Obvious examples are card games, chess, or soccer. If I am to play a card to a trick, then it depends on the cards played by my playing partners whether or not I win the trick. Whether my move in chess leads to a win usually depends on the subsequent moves of my opponent. Whether I should pass the ball to this or that team member depends not in the least on my expectations about whether or not he will pass it on to a player in an even more favourable position. Whether or not my utterance is successful depends on how it is taken up by its addressee and the overall purpose of the current conversation. This provides the basis for applications of game theory in pragmatics.

Game theory has a prescriptive and a descriptive aspect. It can tell us how we should behave in a game in order to produce optimal results, or it can be seen as a theory that describes how agents actually behave in a game. In this book, the latter interpretation of game theory is of interest. The authors of this volume will explore game theory as a framework for describing the use of language.

1.1 Decisions

At the heart of every game theoretic problem there lies a decision problem: one or more players have to choose between several actions. Their choice is governed by their preferences over expected outcomes. If someone is offered a cherry and a strawberry but can only take one of them, then if he prefers the strawberry over the cherry, he will take the strawberry. This is not a prescription. It is an explication of the semantics of the word *preference*. If I can choose between actions a_1 and a_2, and prefer the outcome s_1 of a_1 over s_2 of a_2, then it is the very meaning of the word *preference* that I choose action a_1. In general, one can distinguish between decision making under *certainty*, *risk* and *uncertainty*. A decision is made under *certainty* if the decision maker knows for each action which, outcome it will lead to. The cherry and strawberry example is such a case. A decision is made under *risk* if each action leads to a set of possible outcomes, where each outcome occurs with a certain probability. The decision maker knows these probabilities, or behaves as if he knew them. A decision is made under *uncertainty* if no probabilities for the outcomes are known to the decision maker, and where not even reasonable assumptions can be made about such probabilities. We consider here only decision making under certainty or risk, as does the majority of literature on decision theory.

Decision under risk

Before we enter into game theory proper we want to say more about decision under risk. Decision theory found interesting applications in pragmatics, and its ideas and concepts are fundamental for game theory. The decision maker may be uncertain about the outcomes of his actions because he has only limited information about the true state of the world. If Adam has to decide in the morning whether or not to take an umbrella with him, this depends on whether or not he believes that it will rain that day. He will not know this but will have some expectations about it. These expectations can be represented by probabilities, and Adam's information state by a *probability space*.

We identify a proposition A with sets of possible worlds. In probability theory they are called *events*; but we will stick here to the more familiar terminology from possible worlds semantics. If a person is convinced that A is true, then we assign probability 1 to it, and 0 if he thinks that it can not be true. If there are two propositions A and B that cannot be true at the same time, e.g. that the sky is sunny and that the sky is cloudy, then the probability of A *or* B is just the sum of the probability of A and the probability of B. The latter property is generalised in the following definition to arbitrary countable sequences of pairwise incompatible propositions.

Let Ω be a countable set that collects all possible states of the world. P is a *probability distribution* over Ω if P maps all subsets of Ω to the interval $[0, 1]$ such that:

1 $P(\Omega) = 1$;

2 $P(\sum_{j\in J} A_j) = \sum_{j\in J} P(A_j)$ for each family $(A_j)_{j\in J}$ of countably many pairwise disjoint sets. The sum $\sum_{j\in J} A_j$ here denotes the (disjoint) union of the sets A_j.

We call (Ω, P) a (countable) *probability space*. The restriction to countable Ω's simplifies the mathematics a lot. It follows e.g. that there is a subset $S \subseteq \Omega$ such that $P(\{v\}) > 0$ for each $v \in S$ and $P(A) = \sum_{v \in A \cap S} P(\{v\})$ for all $A \subseteq \Omega$, i.e it follows that P is a *count* measure. For $P(\{v\})$ we write simply $P(v)$.

If (Ω, P) describes the information state of a decision maker, what does his new information state look like if he learns a fact E? Adam may look out of the window and see that the sky is cloudy, or he may consult a barometer and see that it is rising. E would collect all worlds where the sky is cloudy, or, in the second scenario, where the barometer rises. If neither fact contradicts what Adam previously believed, then his probabilities for both sets must be greater than zero. Whatever proposition E represents, how does *learning* E affect (Ω, P)? In probability theory this is modelled by *conditional probabilities*. In learning theory, these are known as *Bayesian updates*. Let H be any proposition, e.g. the proposition that it will rain, i.e. H collects all possible worlds in Ω where it rains at some time of the day. The probability of H *given* E, written $P(H|E)$, is defined by:

$$P(H|E) := P(H \cap E)/P(E) \text{ for } P(E) \neq 0. \tag{1.1}$$

In particular, it is $P(v|A) = P(v)/P(A)$ for $v \in A \neq \emptyset$. E.g. before Adam looked out of the window he may have assigned to the proposition $(E \cap H)$ that it is cloudy *and* that it rains a probability of $\frac{1}{3}$ and to the proposition (E) that it is cloudy a probability of $\frac{1}{2}$. Then (1.1) tells us that, *after* observing that the sky is cloudy, Adam assigns probability $\frac{1}{3} : \frac{1}{2} = \frac{2}{3}$ to the proposition that it will rain. Bayesian updates are widely used as a model for learning. P is often said to represent the *prior* beliefs, and P^+ defined by $P^+(A) = P(A|E)$ the *posterior* beliefs.

As an illustration we want to show how this learning model can be applied in Gricean pragmatics for explicating the notion of *relevance*. We discuss two approaches. The first one measures relevance in terms of the amount of information carried by an utterance and is due to Arthur Merin (Merin 1999b). The second approach introduces a measure that is based on expected utilities and is used by Prashant Parikh (Parikh 1992, Parikh 2001), Rohit Parikh (Parikh 1994) and Robert van Rooij (van Rooij 2003b).

The fact that the barometer is rising (E) provides evidence that the weather is becoming sunny. We can see the situation as a competition between two hypotheses: (H) *The weather will be sunny*, and $(\overline{H})$ *The weather will be rainy*. For simplicity we may assume that H and $\overline{H}$ are mutually exclusive and cover all possibilities. E, the rising of the barometer, does not necessarily imply that H, but our expectations that the weather will be sunny are much higher after learning E than before. Let P represent the given expectations before learning E, i.e. P is a probability distribution over possible states of the world. Let P^+ represent the expectations obtained from epistemic context (Ω, P) when E, and nothing but E, is learned. Modeling learning by conditional probabilities as above, we find that $P^+(H) = P(H|E)$, where we have to assume that $P(E) \neq 0$, i.e. we can only learn something that doesn't contradict our previous beliefs.

Our next goal is to introduce a measure for the *relevance* of E for answering the question whether H or $\overline{H}$ is true. Measures of relevance have been extensively studied in statistical decision theory (Pratt et al. 1995). There exist many different explications of the notion of *relevance* which are not equivalent with each other. We choose here Good's notion of relevance (Good 1950). It was first used by Arthur Merin (Merin 1999b), one of the pioneers of game theoretic pragmatics, in order to get a precise formulation of Grice's Maxim of Relevance.[1]

If we know $P(H|E)$, then we can calculate the reverse, the probability of E given H, $P(E|H)$, by *Bayes' rule*:

$$P(E|H) = P(H|E) \times P(E)/P(H). \tag{1.2}$$

With this rule we get:

$$P^+(H) = P(H|E) = P(H) \times (P(E|H)/P(E)). \tag{1.3}$$

$\overline{H}$ denotes the complement of H. Learning E influences our beliefs about $\overline{H}$ in the same way as it influences our beliefs about H: $P^+(\overline{H}) = P(\overline{H}|E)$. This leads us to:

$$\frac{P^+(H)}{P^+(\overline{H})} = \frac{P(H|E)}{P(\overline{H}|E)} = \frac{P(H)}{P(\overline{H})} \times \frac{P(E|H)}{P(E|\overline{H})}. \tag{1.4}$$

Probabilities are non-negative by definition. In addition we assume that all probabilities in this equation are positive, i.e., strictly greater than 0. This allows us to apply a mathematical trick and build the log of both sides of this equation. As the logarithm is strictly monotone it follows that (1.4) is true exactly iff

$$\log(P^+(H)/P^+(\overline{H})) = \log(P(H)/P(\overline{H})) + \log(P(E|H)/P(E|\overline{H})). \tag{1.5}$$

We used here the fact that $\log(x \times y) = \log x + \log y$. Furthermore we know that $\log x = 0$ iff $x = 1$. This means that we can use the term $r_H(E) := \log(P(E|H)/P(E|\overline{H}))$ as a measure for the ability of E to make us believe H. If it is positive, E favors H, if it is negative, then E favors $\overline{H}$. In a competitive situation where a speaker wants to convince his addressee of some proposition H it is reasonable to call a fact E more relevant the more evidence it provides for H. Merin calls $r_H(E)$ also the *argumentative force* of E.[2]

Whether or not this is a good measure of relevance in general depends on the overall character of communication. Merin sees the aim to convince our communication partner of something as the primary purpose of conversation. If Adam has an interview for a job he wants to get, then his goal is to convince the interviewer that he is the right person for it (H). Whatever he says is more *relevant* the more it favors H and disfavors the opposite proposition. We could see this situation as a battle between two agents, H and $\overline{H}$, where assertions E are the possible moves, and where $\log(P(E|H)/P(E|\overline{H}))$ measures the win for H and the loss for $\overline{H}$. Using the terminology that we will introduce in subsubection 1.2.1, we can say that this is a *zero-sum* game between H and $\overline{H}$.

We want to elaborate a little more on this example. The basis for Merin's proposal lies in the assumption that the main purpose of communication is to provide evidence that helps one decide whether a proposition H or its opposite is true. Hence, it works fine as long as we concentrate on yes-no questions or situations where one person tries to convince an addressee of the truth of some hypothesis. In general decision problems, the decision maker has to decide between different actions. Hence, the preferences over outcomes of these actions become important. It is not surprising that we find examples where a measure of relevance based on pure information becomes inadequate. Imagine that Ω consists of four worlds $\{v_1, \ldots, v_4\}$ of equal probability and that the decision maker has to decide between two actions a_1 and a_2. Suppose that she prefers a_1 in v_1 and v_2 and a_2 in v_3 and v_4 but that the value she assigns to a_1 in v_1 is very large compared to the other cases. If the decision maker learns $E = \{v_2, v_3\}$, then, using Merin's measure, this turns out to be irrelevant for deciding whether it is true that it is better to perform a_1 (i.e. $H = \{v_1, v_2\}$), or a_2 (i.e. $\overline{H} = \{v_3, v_4\}$) because $\log(P(E|H)/P(E|\overline{H})) = 0$. But, intuitively, it is relevant for the decision maker if she learns that the most favoured situation v_1 cannot be the case.

Let us return to the job interview example, and turn from Adam the interviewee to the interviewer. Let's call her Eve. From Eve's perspective the situation can be seen as a decision problem. She has to decide between two actions, *employ Adam* (a_1) or *not employ Adam* (a_2). Depending on the abilities of Adam these actions will be differently successful. The abilities are part of the various possible worlds in Ω. We can represent the success of the actions as seen by Eve by her preferences over their outcomes. We assume here that we can represent these preferences by a (von Neumann-Morgenstern) *utility measure*, or *payoff* function U that maps pairs of worlds and actions to real numbers. How does U have to be interpreted? If v is a world in Ω, then an equation like $U(v, a_1) < U(v, a_2)$ says that the decision maker prefers the outcome of action a_2 in v over the outcome of a_1 in v. $U(v, a_1)$ and $U(v, a_2)$ are real numbers, hence their difference and sum are defined. In utility theory, it is generally assumed that utility measures are unique up to *linear rescaling*, i.e. if $U(v, a) = r \times U'(v, a) + t$ for some real numbers $r > 0$ and t and all v, a, then U and U' represent the same preferences. If Eve values employing an experienced specialist twice as much as employing a trained and able novice, and she values employing an able novice as positively as she values employing an inexperienced university graduate negatively, then this can be modeled by assigning value 2 in the first case, value 1 in the second and value -1 in the third. But it could equally well be modeled by assigning 5 in the first, 3 in the second and -1 in the third case. Putting these parts together we find that we can represent Eve's decision problem by a structure $((\Omega, P), \mathcal{A}, U)$ where:

1 (Ω, P) is a probability space representing Eve's information about the world;

2 $\mathcal{A}$ is a set of actions;

3 $U : \Omega \times \mathcal{A} \longrightarrow \mathbf{R}$ is a utility measure.

In decision theory it is further assumed that decision makers optimize *expected utilities*. Let $a \in \mathcal{A}$ be an action. The *expected utility* of a is defined by:

$$EU(a) = \sum_{v \in \Omega} P(v) \times U(v, a) \tag{1.6}$$

Optimizing expected utilities means that a decision maker will choose an action a only if $EU(a) = \max_{b \in \mathcal{A}} EU(b)$. Let's assume that Eve assigns a probability of $p = \frac{3}{4}$ to the proposition that Adam is an inexperienced novice, but gives a probability of $1 - p = \frac{1}{4}$ to the proposition that he has some training. We further assume that she assigns value 1 to employing him in the first case, and value -1 to employing him in the second case. Furthermore, we assume that she values the state where she employs no candidate with 0. Then her expected utilities for employing and not employing him are $EU(a_1) = \frac{3}{4} \times (-1) + \frac{1}{4} \times 1 = -\frac{1}{2}$ and $EU(a_2) = 0$ respectively. Hence she should not employ Adam.

This may represent the situation before the interview starts. Now Adam tells Eve that he did an internship in a company X specialized in a similar field. This will change Eve's expectations about Adam's experience, and thereby her expected utilities for employing or not employing him. Using the ideas presented before, we can calculate the expected utility of an action *a after learning A* by:

$$EU(a|A) = \sum_{v \in \Omega} P(v|A) \times U(v,a); \tag{1.7}$$

where $P(v|A)$ denotes again the conditional probability of v given A. If Eve thinks that the probability that Adam is experienced increases to $\frac{3}{4}$ if he did an internship (A), then the expected utility of employing him now rises to $EU(a_1|A) = \frac{1}{2}$. Hence, Adam was convincing and will be employed. A number of people (P. Parikh, R. Parikh, R. van Rooij) proposed measuring the *relevance* of a proposition A in terms of how it influences a decision problem that underlies the current communication. Several possible ways to measure this influence have been proposed. One heuristic is to say that information A is relevant if and only if it makes a decision maker choose a different action from before, and it is more relevant the more it increases the expected utility. This is captured by the following measure of *utility value* of A:

$$UV(A) = \max_{a \in \mathcal{A}} EU(a|A) - EU(a^*|A). \tag{1.8}$$

a^* denotes here the action the decision maker had chosen before learning A — in our example this would have been a_2, not employing Adam. The expected utility value can only be positive in this case. If Eve had already a preference to employ Adam, then this measure would tell us that there is no relevant information that Adam could bring forward. So, another heuristic says that information is more relevant the more it increases expectations. This is captured by the following measure:

$$UV'(A) = \max_{a \in \mathcal{A}} EU(a|A) - \max_{a \in \mathcal{A}} EU(a). \tag{1.9}$$

If we put the right side in absolutes, then it means that information is the more relevant the more it changes expectations. This would capture cases where Adam could only say things that diminish Eve's hopes.

$$UV''(A) = |\max_{a \in \mathcal{A}} EU(a|A) - \max_{a \in \mathcal{A}} EU(a)|. \tag{1.10}$$

Obviously, Adam should convince Eve that he is experienced. Following Merin we could say that arguments are more relevant for Adam if they favor this hypothesis and disfavor the opposite. If Adam uses the utility-based measure of relevance, then he should choose arguments that make Eve believe that the expected utility after employing him is higher than that after not employing him. Given our scenario, this is equivalent with choosing arguments that favor the thesis that he is experienced. Hence, we see that for special cases both measures of relevance may coincide.

We want to conclude this section about decision theory with a classical example of Grice's. In this example there is no obvious hypothesis for which the provider of information could argue. Nevertheless, we can explain the relevance of his statement by a criterion based on the maximization of expected utilities.

A and B are planning their summer holidays in France. A has an open map in front of him. They would like to visit C, an old friend of B. So A asks B: *Where does C live?* B answers: *Somewhere in the south of France*. We are not concerned here with the question of how the implicature '*B does not know where C lives*' arises but with the question why B's answer is relevant. In Merin's model, there must be an hypothesis H that B argues for. But it is not immediately clear what this hypothesis H should be. We can model the situation as a decision problem where Ω contains a world for each sentence *C lives in* x, where x ranges over cities in France and where each of these worlds is equally possible. $\mathcal{A}$ contains all actions a_x of *going to* x, and U measures the respective utilities with $U(v, a) = 1$ if a leads to success in v and $U(v, a) = 0$ if not. Let E be the set of all worlds where C lives in the south of France. Calculating the expected utilities $EU(a_x|E)$ and $EU(a_x)$ for an arbitrary city x in the south of France would show that E increases the expected payoff of performing a_x. Hence, if B has no more specific information about where C lives, then a criterion that measures *relevance* according to whether or not it increases expected utilities would predict that E is the most relevant answer that B could give.

1.2 Games

What distinguishes game theory from decision theory proper is the fact that decisions have to be made with respect to the decisions of other players. We start this section with some fundamental classifications of games and introduce the normal form. We look then at one example in more detail, the *prisoners' dilemma*. In section 1.2.3 we present the most fundamental solution concepts of game theory, especially the concept of a *Nash equilibrium*. Finally, we introduce the extensive form. The latter is more suitable for sequential games, a type of game in terms of which communication is studied a lot.

1.2.1 Strategic Games and the Normal Form

There exist several important different classifications of games which are widely referred to in game theoretic literature. We provide here a short overview.

A first elementary distinction concerns that between *static* and *dynamic games*. In a static game, every player performs only one action, and all actions are performed *simultaneously*. In a dynamic game there is at least one possibility of performing several actions in *sequence*.

Furthermore, one distinguishes between *cooperative* and *non-cooperative* games. In a cooperative game, players are free to make binding agreements in preplay communications. Especially, this means that players can form *coalitions*. In *non-cooperative* games no binding agreements are possible and each player plays for himself. In our discussion of the prisoners' dilemma we will see how the ability to make binding agreements can dramatically change the character and solutions of a game. But, except for this one illustrative example, we will be concerned with non-cooperative games only.

There are two standard representations for games: the *normal form* and the *extensive* form. Our introduction will follow this distinction. The major part concentrates on static games in normal form, which we will introduce in this section. We introduce the extensive form together with dynamic games in section 1.2.4.

Games are played by *players*. Hence, in a description of a game we must find a set of players, i.e. the people who choose actions and have preferences over outcomes. This implies that *actions* and *preferences* must be represented too in our game models. Let $N = \{1, \ldots, n\}$ denote the set of players. Then we assume that for each player there is a set $\mathcal{A}_i$ that collects all actions, or moves, that can be chosen by him. We call $\mathcal{A}_i$ player i's *action set*. An *action combination*, or action *profile*, is a n–tuple $(a_1, \ldots, a_n)$ of actions where each $a_i \in \mathcal{A}_i$. The assumption is that they are performed simultaneously. In general, we distinguish *strategies* from actions. This becomes important when we consider *dynamic* or *sequential* games. Strategies tell players what to do in each situation in a game given their background knowledge and are modeled by functions from sequences of previous events (histories) into action sets. In a *static* game, i.e. a game where every player makes only one move, these two notions coincide. We will use the expressions *strategy sets* and *strategy combinations*, or strategy *profiles*, in this context too, although strategies are only actions.

Players and action sets define what is feasible in static games. The preferences of players are defined over action or strategy profiles. We can represent them either by a binary relation $\preceq_i$, $i = 1, \ldots, n$, between profiles, or by payoff functions u_i mapping profiles to real numbers. If $(s'_1, \ldots, s'_n) \prec_i (s_1, \ldots, s_n)$ or $u_i(s'_1, \ldots, s'_n) < u_i(s_1, \ldots, s_n)$, then player i prefers strategy profile $(s_1, \ldots, s_n)$ being played over strategy profile $(s'_1, \ldots, s'_n)$ being played. We can collect the individual u_i's together in payoff profiles $(u_1, \ldots, u_n)$ and define the *payoff function* U of a game as a function that maps all action or strategy profiles to payoff profiles. A static game can be represented by a *payoff-matrix*. In the case of two-player games with two possible actions for each player it has the form given in Table 1.1.

Table 1.1: Payoff-matrix of a two-player game

	b_1	b_2
a_1	$(u_1(a_1, b_1)\ ;\ u_2(a_1, b_1))$	$(u_1(a_1, b_2)\ ;\ u_2(a_1, b_2))$
a_2	$(u_1(a_2, b_1)\ ;\ u_2(a_2, b_1))$	$(u_1(a_2, b_2)\ ;\ u_2(a_2, b_2))$

One player is called *row player*, he chooses between actions a_1 and a_2; the other player is called *column player*, he chooses between actions b_1 and b_2. We identify the row player with player 1, and the column player with player 2. The action set $\mathcal{A}_1$ of player 1 is then $\{a_1, a_2\}$, and that for player 2 is $\mathcal{A}_2 = \{b_1, b_2\}$. $u_i(a_k, b_l)$ is the payoff for player i for action profile (a_k, b_l). It is assumed that two payoff functions U and U' are equivalent, i.e. represent the same preferences, if there is an $r > 0$ and a t such that for all $i = 1, \dots, n$ and $a \in \mathcal{A}$: $ru_i(a) + t = u'_i(a)$.

Hence, in the class of games we introduced here the players' payoffs depend only on the actions chosen, and not on the state of the environment. In the next section we discuss an example. Putting things together, we define a *strategic game* as a structure $(N, (\mathcal{A}_i)_{i \in N}, U)$ such that:

1 $N = \{1, \dots, n\}$ the (finite) set of players $1, \dots, n$;

2 $\mathcal{A}_i$ is a non-empty set of actions for each player $i \in N$; $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$ is the set of all action profiles.

3 $U : \mathcal{A} \longrightarrow \mathbf{R}^n$ is a payoff function which maps each action profile $(a_1, \dots, a_n) \in \mathcal{A}$ to an n–tuple of real numbers $(u_1, \dots, u_n)$, i.e. $(u_1, \dots, u_n)$ is the the payoff profile of players $1, \dots, n$ for action profile $(a_1, \dots, a_n)$.

The following notation is very common in connection with profiles: if $s = (s_1, \dots, s_n)$ is a given strategy profile, action profile, or payoff profile etc., then s_{-i} denotes the profile $(s_1, \dots, s_{i-1}, s_{i+1}, \dots, s_n)$, $1 \leq i \leq n$; i.e. s_{-i} is the profile of length $n - 1$ that we get if we eliminate player i's strategy, action, payoff etc. (s'_i, s_{-i}) then denotes the profile where we have replaced s_i in the original profile s by s'_i.

We can classify strategic games according to how much the payoff functions of the players resemble each other. One extreme are *zero-sum* games, or *strictly competitive* games; the other extreme are games of *pure coordination*. In a zero-sum game the payoffs of the players sum up to zero for each strategy profile. This means, that if one player wins a certain amount, then the other players lose it. These games are strictly competitive and if they are played by two persons we could justly call them opponents. A game of pure coordination is exactly the opposite extreme where the payoffs of all players are identical for each action profile. If one player wins something then the other player wins the same amount, and if one player loses then the other one loses too. We really could call them partners. Zero-sum games and games of pure coordination are two ends on a scale ranging from pure conflict to its opposite. In between are cases where interests partially overlap and partially conflict.

In the last section we saw an example of a zero-sum game. In Merin's approach to pragmatics, the aim to convince one's conversational partner of some hypothesis H is the basic dialogue situation. This was modeled by a zero-sum game where the players are the hypotheses H and $\overline{H}$, the complement of H, the moves are propositions E, and the payoffs are defined by the relevance of E for the respective hypotheses. If E favors H, then it disfavors $\overline{H}$ the same amount, and vice versa. Games of pure coordination are fundamental if we look, following David Lewis, at language as a *convention*.

1.2.2 The Prisoners' Dilemma and Strict Domination

That a decision is not made under certainty does not necessarily imply that we have to calculate expectations expressed in terms of probabilities. Often we can do without them. Suppose there are two supermarkets, both sell exactly the same goods, and it takes the same effort to go shopping at one as to the other. I like to buy vanilla ice cream, but if there is no vanilla ice cream, I want strawberry ice cream, or ice cream with orange flavor. And I want to buy it as cheaply as possible. I know that one supermarket A sells everything at a lower price than the other supermarket B. Hence, whatever the state of the world, whatever sorts of ice cream they sell, I can never be better off if I go to supermarket B. *To be better off* means here to have a preference for the outcome resulting from one action, shopping at A, over the outcome resulting from the other action, shopping at B. Now, assume that I know in addition that at least one of my favorite sorts of ice cream is in store. So what to do? If I have to decide between actions a_1 and a_2, and whatever the state of the world, I strictly prefer the outcome of a_1 over that of a_2, then I will choose action a_1. We say that an action a_1 *strictly dominates* an action a_2 if in all possible courses of events the results from performing a_1 are strictly preferred over that of a_2. It then amounts to a tautology to say that an agent will never choose the strictly dominated action. This criterion may tell us what a decision maker will do although he does not know the results of his actions with certainty, and although his expectations about these results are unknown.

This example is an example of a pure decision problem, i.e. a problem where the outcome of the choice of action solely depends on the state of the world and not on the decisions of other players. It is straightforward to generalize the last principle of strict domination to proper game situations: if I have to decide between actions a_1 and a_2, and, whatever actions the other players choose, the outcomes resulting from a_1 are always strictly preferred over the outcomes of a_2, then I will choose action a_1. Again, this is meant as a tautology. As a consequence, if we study a decision problem in a game and ask which action will be chosen by the players, then we can eliminate all strictly dominated actions without losing any of the reasonable candidates.

The *prisoners' dilemma* is one of the most discussed problems in game theory. It is a standard example illustrating the principle of elimination of strictly dominated actions. One version of the story runs as follows: the police arrest two gangsters for a crime they committed together, but lack sufficient evidence. Only if they confess can the police convict them for this crime. Hence, the police separate them, so that they can't communicate, and offer each of them a bargain: if he confesses, and the other one doesn't, then he will be released but his companion will be sentenced to the maximum penalty. If both confess, then they still will be imprisoned but only for a considerably reduced time. If neither of them confesses, then the police can convict them only for a minor tax fraud. This will be done for sure and they both will receive a minor penalty. The exact numbers are irrelevant but they help to make examples more intuitive. So, let's say that the maximal penalty is 10 years, the reduced penalty, in the case where both confess, is 8 years, the tax fraud is punished with 2 years, and if they are released they are imprisoned for 0 years. The police inform both criminals that they offer this bargain to each of them, and that they are both informed about this. Graphically we can represent the decision situation of the two prisoners as in Table 1.2.

Table 1.2: The prisoners' dilemma

	c	d
c	$(-2\,;\ -2)$	$(-10\,;\ 0)$
d	$(0\,;\ -10)$	$(-8\,;\ -8)$

Each of the two players has to choose between cooperating and non-cooperating with his companion. We denote these actions by c (cooperate) and d (defect). If a prisoner defects, then he confesses; if he cooperates, then he keeps silent. One prisoner chooses between columns, i.e. he is the *column player*, the other between rows, i.e. he is the *row player*. This payoff-matrix tells us e.g. that column player will be sentenced to 0 years if he defects, and if at the same time row player cooperates, the row player will be sentenced to 10 years.

It is easy to see that for both players action d strictly dominates action c. Whatever the other player chooses, he will always prefer the outcome where he himself had performed d. Hence, after elimination of strictly dominated actions, only the pair (d, d) remains as a possible choice, and hence both will confess and be sentenced to 8 years.

The prisoners' dilemma is an instructive example not the least because it easily gives rise to confusion. It seems to lead into a paradox: if both players strictly follow their preferences, then they are led to perform actions with results that are much disliked by both. But to say that somebody follows his preferences is no more than a tautology, so the players cannot do anything else but strictly follow them. The question may arise whether the principle of eliminating strictly dominated actions isn't too simple minded. It is necessary to make clear what the game theoretic model describes and what it doesn't describe.

As mentioned, the model has to be understood as a descriptive model, not as a prescriptive one. Hence, it does not advise us to follow only our narrowly defined short-term advantages and disregard all needs and feelings of our companions. It just says if the preferences are such as stated in the model, then a rational player will act in this and that way.

It makes a crucial difference whether the prisoners' dilemma is played only once or whether it is played again and again. In the repeated prisoners' dilemma there is a chance that we meet the same person several times, and non-cooperative behavior can be punished in future encounters. And, indeed, it can be shown that there are many more strategies that rational players can choose when we considered the infinitely repeated prisoners' dilemma.

The model assumes that the preferences of the players are just as stated by the payoff matrix. This means that the only thing the prisoners are interested in is how long they will be imprisoned. They are not interested in the fates of each other. Again this is not a prescription but a description of a certain type of situation. If we consider a scenario where the prisoners feel affection for each other, then this has to be represented in the payoff matrix. In an extreme case where one criminal cares as much for the other one as for himself, his payoffs may be just the negative sum of both sentences. In this case the model would predict that the compassionate prisoner cooperates, and this behavior is as rational as the defecting behavior in the first scenario. The corresponding payoff-matrix is given in Table 1.3.

Table 1.3: The prisoners' dilemma with a compassionate row player

	c	d
c	$(-4\,;\,-2)$	$(-10\,;\,0)$
d	$(-10\,;\,-10)$	$(-16\,;\,-8)$

But still, there may remain a gnawing doubt. Let's assume that we play the prisoners' dilemma only once, and let's assume that the preferences are exactly the same as those stated in its payoff matrix, isn't it simply better to cooperate? Even if we were convinced that the only rational choice is to defect, doesn't this example show that it is sometimes better to be irrational? – and thereby challenge a central game theoretic assumption about rational behavior?

First, it has to be noted that the prisoners' dilemma does not show that it is better to deviate from defection unilaterally. As the example with the compassionate row player shows, this will simply mean to go to jail for 10 years. But, of course, it would be *better* for both to cooperate simultaneously. This simply follows from the payoff matrix; but this does not imply that the principle of elimination of dominated actions is *irrational*. The source for the doubt lies in the observation that if they were bound by an agreement or by moral obligations, then they would be better off following it even if this course of action contradicts their own preferences. The point is that in the setting we considered for the one-shot prisoners' dilemma there is no room for agreements or moral obligations. If we add them, then we get a different game, and for *this* game we can indeed find that it is rational to make binding contracts because they lead to better payoffs, even if the payoffs are defined exactly as in the original situation.

Let's assume that the two criminals have a possibility to agree never to betray each other if they get imprisoned. Here is a straightforward model for this situation: we add two actions, a and $-a$. In the beginning both players have to decide whether they play a or $-a$. If both play a, then they make an *agreement* that they both play c, *cooperate*, in the prisoners' dilemma. If one of them doesn't play a, then no agreement is formed and hence nothing changes and both can cooperate or defect afterwards. An agreement is binding; i.e. it is impossible to do anything that is not in accordance with it. Hence, it has the effect of reducing the set of possible actions open to the players. Call the two players A and B. We depict this game as in Figure 1.1.

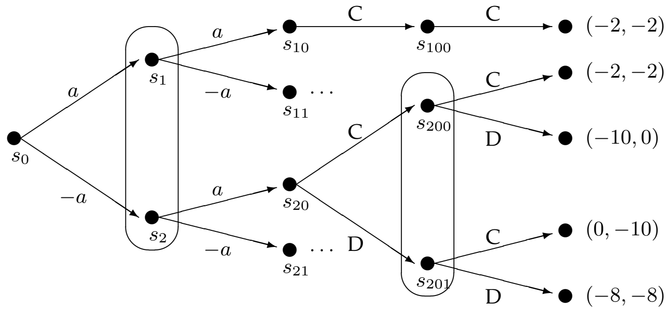


Figure 1.1

In this representation we depict the players' moves one after the other. First A chooses between actions a and $-a$ in situation s_0, which leads to s_1 or s_2 respectively. Then B makes his choice from a and $-a$. The oval around s_1 and s_2 means that B cannot distinguish between the two situations, i.e. he does not know whether A played a or $-a$. Hence, although the actions of A and B are ordered sequentially, the graph covers also the case where both decide simultaneously. We will introduce this form of representation, the extensive form, in section 1.2.4. After their initial choice, both have to play the prisoners' dilemma. As in s_{11}, s_{20} and s_{21} no agreement is reached, both play the ordinary prisoners' dilemma as considered before. We depicted the associated game tree only once after s_{20}. It is identical for all three situations. In situation s_{10} they reached an agreement. So their possibilities in the prisoner's situation are limited to cooperation. This means that they have only one choice to act. In the end we find the payoffs for each course of events. The first value is the payoff for player A and the second for player B.

Which actions will rational players choose? As the situations in s_{11}, s_{20} and s_{21} are those of the prisoners' dilemma, it follows that both will play defect. Hence, their payoffs for all these situations will be $(-8,-8)$. If they are in s_{10} their payoff will be $(-2,-2)$. For their choice between a and $-a$ in the initial situation this means that the game tree can be simplified as depicted in Figure 1.2.

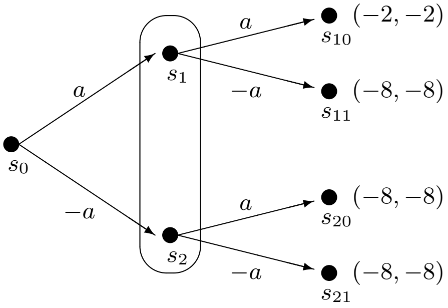


Figure 1.2

If A has played a, then B has a preference to play a too; if A played $-a$, then B has no preferences for one over the other. The situation for A is symmetrical. We can represent this game by a payoff matrix in Table 1.4.

Table 1.4

	a	$-a$
a	$(-2\,;\,-2)$	$(-8\,;\,-8)$
$-a$	$(-8\,;\,-8)$	$(-8\,;\,-8)$

The principle of elimination of strictly dominated actions does not help us here because a is not preferred by the players for every choice that the other player can make. This example motivates a more general principle, that of the elimination of weakly dominated actions: if a player has a choice between actions $a_1, \ldots, a_n$ such that (1) there is a possibility where a_1 leads to a preferred outcome, and (2) there is no possibility where its outcome is dispreferred, then the player will choose a_1. This principle is already more than a pure explication of the meaning of *prefer*. It presupposes some deliberation on the side of the player. If we apply this principle in the situation represented by the last payoff matrix, we see that (a, a) is the only possible choice for the two agents.

What does this example show us? We saw that it would be better for both of the prisoners if they could manage to cooperate in the prisoners' dilemma, and this threw some doubt on the principle of elimination of strongly dominated actions. As we see now, if there is a device that forces them to cooperate, then it is rational for them to use it. But to restrict one's own freedom and force oneself to make a certain choice is something different from behaving irrationally. To choose *cooperate* in the original prisoners' dilemma means to make an irrational choice; which is in that case tantamount to preferring the dispreferred. The making of binding agreements is part of the actions the players can perform. If we want to study its effects on finding optimal behavior, then it has to be represented in the game structure. Doubts about the principle of elimination of strongly preferred actions arise when we put the possibility of deliberately restricting our freedom into the criterion of rationality. This shows that we have to be careful about what we model by what in a game.

We can distinguish three aspects of a decision problem that have to be represented in a model. An agent has to consider: (1) what is feasible; (2) what is desirable; (3) what he knows. Hence, we have to represent in each decision situation what a player can do, what his preferences are and what he knows. In our game tree we can read off this information as follows: in situation s_{20} there are two possible actions player B can choose, and they lead to situations s_{200} and s_{201} respectively. This defines the *feasible*. His preferences are defined over final outcomes of courses of events. They are given by his payoffs in the final situations. He knows that he is in situation s_{20}. If he couldn't distinguish between this situation and another, then this could be indicated by an oval containing both situations. This is called his *information set*. But this represents only his special knowledge about the specific game situation. In addition to this, it is usual to assume that players know the overall game structure, i.e. they know which moves are possible in which situations by each player, know their consequences, know each other's preferences, and each other's knowledge about each other's moves. More precisely, the latter properties have to be *common* or *mutual knowledge* . A proposition φ is mutually known if every player knows φ, if every player knows that every player knows that φ, if every player knows that every player knows that every player knows that φ, etc. Hence, this imposes a very strong assumption about the players' ability to reason about each other. This is a characteristic of classical Game Theory . In a later section we will see an interpretation of Game Theory that imposes much weaker constraints on the rationality of players: evolutionary game theory .

The components introduced so far, the possible moves, the desires, and the players' knowledge provide a description of a game but they still do not provide an answer to the question what an agent will or should do. What is needed is a criterion for selecting an action. In the prisoners' dilemma we saw one such criterion, the principle of elimination of strongly dominated strategies. If an agent is *rational*, we said, then he cannot choose an action that will always lead to a dispreferred state of affairs. We introduced it as an explication of the meaning of *prefer*. But in general, this principle will not suffice to give us an answer for every game where we expect an answer. We saw already an example where we needed a weaker version of the principle. In the game theoretic literature we find a large number of criteria for what is *rational* to choose. These criteria may depend on the type of game played, the information available to the players, on assumptions on their ability to reason about each other, on the amount of common knowledge. Very often, these criteria are formulated in terms of *equilibrium concepts*.

1.2.3 Strategic games and the Nash equilibrium

Strategic Games without Uncertainty What strategy will a rational player choose in a given strategic game ? We saw one way to answer this question: a player may just eliminate all strictly dominated actions, and hope to find thereby a single possible move that remains, and hence will be chosen by him. We formulate *strict domination* for strategies:

Definition 1 (Strict Domination) *A strategy s_i of player i strictly dominates a strategy s'_i, iff for all profiles s it holds that $(s'_i, s_{-i}) \prec_i (s_i, s_{-i})$.*

For *weak* domination we have to replace $\prec_i$ by $\preceq_i$ and to assume that there is at least one strategy combination by the opponents such that s_i does better than s'_i.

Definition 2 (Weak Domination) *A strategy s_i of player i weakly dominates a strategy s'_i, iff (1) for all profiles s it holds that $(s'_i, s_{-i}) \preceq_i (s_i, s_{-i})$, and (2) there is a profile s such that $(s'_i, s_{-i}) \prec_i (s_i, s_{-i})$.*

Whereas we can see the principle of strict domination as a mere explication of the meaning of *prefer*, the principle of weak domination involves some reasoning on the side of the player. Only if there is a chance that the other players choose the strategies s_{-i}, where s_i is preferred over s'_i, there is a reason to play s_i. In the previous examples we applied the principles of elimination of dominated strategies only once for each player. But in many games we have to apply them several times to arrive at a unique solution. For instance, consider the game in Table 1.5 on the facing page. Intuitively, the combination (r_1, c_1) should turn out as the solution. This is a game with two players. We call them again row player and column player. Each player has the choice between three actions; row player between $\{r_1, r_2, r_3\}$ and column player between $\{c_1, c_2, c_3\}$. Neither c_1, c_2 nor c_3 are dominated, hence our criterion does not tell us what column player will choose. For the row player neither r_1 nor r_2 are dominated; r_1 is better if column player chooses c_1 and c_2, and r_2 is better if he chooses c_3. But we can see that r_2

Table 1.5

	c_1	c_2	c_3
r_1	$(5\ ;\ 5)$	$(4\ ;\ 4)$	$(0\ ;\ 0)$
r_2	$(1\ ;\ 1)$	$(3\ ;\ 3)$	$(2\ ;\ 2)$
r_3	$(0\ ;\ 0)$	$(0\ ;\ 0)$	$(1\ ;\ 1)$

strictly dominates r_3. This means that row player will never choose this action. Now, if we assume that the game structure is common knowledge, i.e. the payoffs are mutually known, and if column player knows about row player that he eliminates strictly dominated action, then column player can infer too that only r_1 and r_2 are possible moves by row player. If we assume that this reasoning is mutually known, then we can eliminate the third row of the payoff matrix. In the reduced matrix c_3 is strictly dominated by c_2, and for the same reasons we can eliminate it from the payoff matrix. In the remaining 2×2 matrix, r_1 strictly dominates r_2. Hence there remains only one choice for row player. This means that the problem of what to choose is solved for him. But if only r_1 is a possible move for row player, then c_1 strictly dominates c_2, and therefore the problem is solved for column player too. It turns out that (r_1, c_1) is their unique choice.

Apart from the fact that we have to apply the principle of elimination of dominated strategies iteratively, there is another important point that is confirmed by this example: that row player and column player arrive at (r_1, c_1) presupposes that they know about each other that they apply this principle and that they are able to work out quite intricate inferences about each other's behavior. Such assumptions about the agents' ability to reason about each other play a major role in the justification of all criteria of rational behavior.

Unfortunately, iterated elimination of dominated strategies can't solve the question how rational players will act for all types of static games. The example in Table 1.6 is known as the *Battle of the Sexes*. There are several

Table 1.6: Battle of the sexes

	b	c
b	$(4\ ;\ 2)$	$(1\ ;\ 1)$
c	$(0\ ;\ 0)$	$(2\ ;\ 4)$

stories told for this game. One of them runs as follows: row player, let's call him Adam, wants to go to a boxing event this evening, and column player, let's call her Eve, to a concert. Both want to go to their events together. Eve would rather go to the boxing event with Adam than going to her concert alone although she doesn't like boxing very much. The same holds for Adam if we reverse the roles of boxing and the concert.

We see that for Adam neither going to the boxing event b dominates going to the concert c, nor the other way round. The same holds for Eve. Hence, the principle of elimination of dominated strategies does not lead to a solution to the question what Adam and Eve will do. Intuitively, Adam and Eve should agree on (b, b) and (c, c) if they want to maximize their payoffs. They should avoid (b, c) and (c, b). What could a justification look like? One way of reasoning proceeds as follows: if Eve thinks that Adam will go to the boxing event, then she has a preference to be there too. However, if Adam knows that Eve goes to the boxing event, then Adam wants to be there too. The same holds for the pair (c, c). If we look at (b, c), then we find that in this case Adam would prefer to play c, as this increases his payoff; or, if Eve knows that Adam plays b, then she would prefer to switch to b. A strategy profile $s = (s_1, \ldots, s_n)$ is called a *Nash equilibrium* if none of the players i has an interest in playing a strategy different from s_i given that the others play s_{-i}.

Definition 3 (Nash Equilibrium) *A strategy profile s is a* (weak) Nash equilibrium *iff for none of the players i there exists a strategy s'_i such that $s \prec_i (s'_i, s_{-i})$, or, equivalently, if for all of i's strategies s'_i it holds that $(s'_i, s_{-i}) \preceq_i s$.*

A Nash equilibrium is *strict* if we can replace the $\preceq_i$ by $\prec_i$ for $s'_i \neq s_i$ in the second characterization. In this case every player has a preference to play s_i if the others play s_{-i}.

There is another characterization, in terms of best responses. A move s_i of player i is a *best response* to a strategy profile s_{-i}. We write $s_i \in BR_i(s_{-i})$, iff

$$u_i(s_i, s_{-i}) = \max_{s'_i \in S_i} u_i(s'_i, s_{-i}). \tag{1.11}$$

A strategy profile s is a Nash equilibrium, iff for all $i = 1, \ldots, n$ s_i is a best response to s_{-i}, i.e. iff $s_i \in BR_i(s_{-i})$. It is *strict* if in addition $BR_i(s_{-i})$ is a singleton set for all i.

Mixed Nash Equilibria We saw that the Battle of the Sexes has two Nash equilibria, (b, b) and (c, c). Once they managed to coordinate on one of them, they have no reason to play something else. But if it is the first time they play this game, or if they have to decide what to play every time anew, then the existence of two equilibria does not give them an answer to their question what to do. Eve may reason as follows: if I go to the boxing event, then the best thing I can get out is 2 pleasure points, but if things go wrong then I get nothing. But if I go to the concert, then the worst thing that can happen is that I get only 1 pleasure point, and if I am lucky I get 4 points. So, if I play safe, then I should avoid the action that carries with it the potentially worst outcome. Hence, I should go to the concert. If Adam reasons in the same way, then he will go to the boxing event, and hence Adam and Eve will always be at their preferred event but never meet and get limited to payoff 1. The strategy that Adam and Eve play here is known as the *minimax strategy*. When von Neumann and Morgenstern wrote their seminal work *On the Theory of Games and Economic Behavior* (1944) they didn't use Nash equilibria as their basic solution concept but the minimax strategy. Obviously, this strategy is reasonable. But we will see that Adam and Eve can do better.

What happens if Adam and Eve *mix* their strategies, i.e. they don't choose a *pure* strategy like *always* playing c or b but play each strategy with a certain *probability*? Let us assume that in the situation of Table 1.1. Adam goes to the concert with probability $\frac{1}{2}$ and to the boxing event with probability $\frac{1}{2}$, and Eve does the same. Then the probability of each of the four possible combination of actions is $\frac{1}{2} \times \frac{1}{2}$. The situation is symmetrical for both players, hence they can both calculate their expected payoffs as follows:

$$\frac{1}{2} \times \frac{1}{2} \times 2 + \frac{1}{2} \times \frac{1}{2} \times 1 + \frac{1}{2} \times \frac{1}{2} \times 0 + \frac{1}{2} \times \frac{1}{2} \times 4 = 1\frac{3}{4}$$

So we see that a simple strategy like flipping a coin before deciding where to go can improve the overall outcome significantly.

What is a Nash equilibrium for games with mixed strategies? If Eve believes that chances are equally high for Adam going to the concert as for Adam going to the boxing match, then she can calculate her expected payoff, or expected utility, of playing c as follows:

$$EU(c) = \frac{1}{2} \times 1 + \frac{1}{2} \times 4 = 2\frac{1}{2}$$

Whereas her expected utility after playing b is

$$EU(b) = \frac{1}{2} \times 2 + \frac{1}{2} \times 1 = 1\frac{1}{2}$$

In general she will find that always playing c is the most advantageous choice for her. A Nash equilibrium is a sequence of choices by each player such that none of the players has an interest to play something different given the choices of the others. Hence, playing b and c both with probabilities $\frac{1}{2}$ can't be a Nash equilibrium. But we will see that there exists one, and, moreover, that there exists one for every finite (two-person) game. Before we introduce this result, let us first state what a strategic game with mixed strategies is.

Let $\Delta(A_i)$ be the set of probability distributions over A_i, i.e. the set of functions P that assign a probability $P(a)$ to each action $a \in A_i$ such that $\sum_{a \in A_i} P(a) = 1$ and $0 \leq P(a) \leq 1$. Each P in $\Delta(A_i)$ corresponds to a *mixed strategy* of agent i. A *mixed strategy profile* then is a sequence $(P_1, \dots, P_n)$ for the set of players $N = \{1, \dots, n\}$. A *pure* strategy corresponds to a mixed strategy P_i where $P_i(a) = 1$ for one action $a \in A_i$ and $P_i(b) = 0$ for all other actions. In our example of the Battle of the Sexes the players' action sets are $\{b, c\}$, i.e. the actions of going to a boxing event and going to a concert. If Adam is player 1 and Eve 2, then their strategy of playing b and c with equal probability corresponds to the strategy profile (P_1, P_2) where $P_i(b) = P_i(c) = \frac{1}{2}, i \in \{1, 2\}$.

We can calculate the *expected utility* of player i given a mixed strategy profile $P = (P_1, \dots, P_n)$ and payoff profile $(u_1, \dots, u_n)$ by:

$$EU_i(P) = \sum_{a \in A_1 \times \ldots \times A_n} P_1(a_1) \times \ldots \times P_n(a_n) \times u_i(a). \tag{1.12}$$

It is assumed that rational players try to maximize their expected utilities, i.e. a player i strictly prefers action a over action b exactly if the expected utility of a is higher than the expected utility of b.

For mixed strategy profiles $P = (P_1, \dots, P_n)$, we use the same notation P_{-i} as for (pure) strategy profiles to denote the profile $(P_1, \dots, P_{i-1}, P_{i+1}, \dots, P_n)$ where we leave out the strategy P_i. (P'_i, P_{-i}) denotes again the profile where we replaced P_i by P'_i.

Definition 4 (Mixed Nash Equilibrium) *A* (weak) mixed Nash equilibrium *is a mixed strategy profile $(P_1, \dots, P_n)$ such that for all $i = 1, \dots, n$ and $P'_i \in \Delta(A_i)$ it holds that $EU_i(P'_i, P_{-i}) \leq EU_i(P)$. A mixed Nash equilibrium is* strict *if we can replace $\leq$ by $<$ in the last condition.*

A standard result states that every finite strategic two-person game has a mixed Nash equilibrium. In the case of our example of the Battle of the Sexes we find that the pair (P_1, P_2) with $P_1(b) = \frac{4}{5}$ and $P_1(c) = \frac{1}{5}$ for Adam and $P_2(b) = \frac{1}{5}$ and $P_2(c) = \frac{4}{5}$ is a mixed Nash equilibrium. If Adam plays P_1, then Eve can play whatever she wants, she never can get a higher payoff than that for playing P_2. In fact, it is the same for all her possible strategies. The analogue holds for Adam if it is known that Eve plays P_2.

That we find a mixed Nash equilibrium for every finite strategic game is of some theoretical interest because there are many games that don't have a pure Nash equilibrium.

Table 1.7

	a	b
a	$(1\,;\,-1)$	$(-1\,;\,1)$
b	$(-1\,;\,1)$	$(1\,;\,-1)$

Consider the game from Table 1.7. This game cannot have a pure Nash equilibrium because whatever one player chooses, because it maximizes his payoff given a choice by the other player, will induce the other player to make a different choice. But it has a unique mixed Nash equilibrium where each player plays each move with probability $\frac{1}{2}$.

There are many refinements of the notion of Nash equilibrium. We introduce here Pareto optimality. We saw that every finite strategic game has a mixed Nash equilibrium . But besides this it may have many more equilibria, hence the criterion to be a Nash equilibrium does normally not suffice for selecting a unique solution for a game. Although this is true, in many cases we can argue that some of these equilibria are better or more reasonable equilibria than others. In the example from Table 1.8 we find two Nash equilibria, (a, a) and (b, b). This means that both are possible solutions to this game but *both* players have an interest to agree on (a, a).

Table 1.8

	a	b
a	$(3\,;\,2)$	$(0\,;\,1)$
b	$(1\,;\,0)$	$(1\,;\,1)$

A Nash equilibrium like (a, a) is called *strongly Pareto optimal*, or *strongly Pareto efficient* . More precisely, *a Nash equilibrium* $s = (s_1, \ldots, s_n)$ is *strongly Pareto optimal*, iff there is no other Nash equilibrium $s' = (s'_1, \ldots, s'_n)$ such that for all $i = 1, \ldots, n$ $u_i(s) < u_i(s')$. I.e. a Nash equilibrium is strongly Pareto optimal, iff there is no other equilibrium where every player is better off. For example, if players can negotiate in advance, then it is reasonable to assume that they will agree on a strongly Pareto optimal Nash equilibrium. There is also a weaker notion of Pareto optimality: a Nash equilibrium is just *Pareto optimal* iff there is no other Nash equilibrium $s' = (s'_1, \ldots, s'_n)$ such that for all $i = 1, \ldots, n$ $u_i(s) \leq u_i(s')$ and for one i $u_i(s) < u_i(s')$.

1.2.4 Games in extensive form

We already saw graphical representations for games in extensive form in Figures 1.1 and 1.2. This form is useful for the representation of dynamic games, i.e. games where there may occur whole sequences of moves by different players e.g. as in chess. The normal form goes together with the matrix representation, and the extensive form with the representation by a *game tree*.

What is a tree? A tree consists of several *nodes*, also called *vertices*, and *edges*. In a game tree an edge is identified with a *move* of some of the players, and the nodes are game situations where one of the players has to choose a move. A tree starts with a distinguished root node, the start situation. If two nodes n_1, n_2 are connected by an edge $n_1 \rightarrow n_2$, then n_1 is a *predecessor* of n_2, and n_2 a *successor* of n_1. Every node has exactly one predecessor, except the root node, which has no predecessor. Every node may have several successors. We call a sequence of nodes $n_1 \rightarrow n_2 \rightarrow \ldots \rightarrow n_k$ a *path* from n_1 to n_k. If n_k has no successor, then we call this sequence a *branch* and n_k an *end node*. In general, a tree may contain also infinite braches, i.e. there may exist an infinite sequence $n_1 \rightarrow n_2 \rightarrow n_3 \rightarrow \ldots$ with no end node. For the following definition of *extensive form game* we assume that there are *no* infinite branches. In a tree, every node, except the root node itself, is connected to the root by a path, and none of the nodes is connected to itself by a path (*no circles*). A tree for an extensive game may look as in Figure 1.3. The numbers $i = 1, 2$ attached to the nodes mean that it is player i's turn to choose a move.

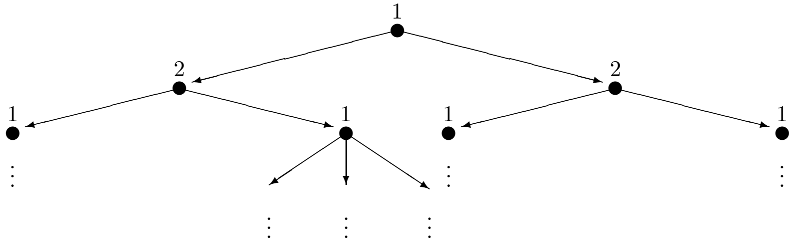


Figure 1.3

In a game like chess all players have *perfect* information, i.e. they know the start situation, each move that either they or their opponents have made, and each others preferences over outcomes.

In Figure 1.1 we saw an example where in some situations players do not know in which situation they are. For every player i we can add to every node n in the tree the set of situations that are indistinguishable for i. This set is called i's *information set*. A typical example where information sets have more than one element is a game where one player starts with a secret move. The second player knows the start situation, which moves player 1 may have chosen, and the possible outcomes. If the move is really secret, then player 2 cannot know in which node of the tree he is when he chooses. Think of player 1 and 2 playing a game where 1 hides a Euro in one hand and 2 has to guess whether it is in 1's left or right hand. If 2 makes the right guess he wins the Euro, otherwise 1 gets it. Figure 1.4 shows a tree for this game. The fact that player 2 cannot distinguish between the situations n, where 1 holds the Euro in his left hand, and n', where he holds it in his right hand, is indicated by the oval around n and n'. What is in general a *game in extensive form*? Let a tree be given in the sense defined above. There must be a set of players $N = \{1, \ldots, n\}$ and a set of moves $\mathcal{A} = \{a_1, \ldots, a_m\}$. Sometimes it is assumed that *nature* is an additional player; then it is denoted by 0. In order to represent a game we need the following information about the tree:

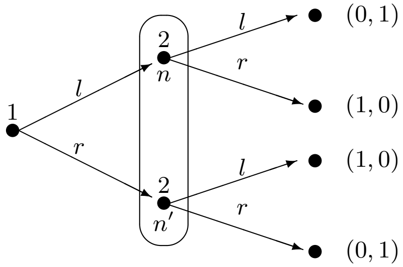


Figure 1.4

1 To each node of the tree we have to assign exactly one player from the set $N \cup \{0\}$. If player i is assigned to a node, then this means that it is player i's turn to make a move in that situation.

2 Each edge has to be labelled by a move from $\mathcal{A}$. If an edge $n_1 \rightarrow n_2$ is labelled with action a and node n_1 is assigned to player i, then this means that playing a by i in situation n_1 leads to n_2.

3 If a node n is assigned to player i, then we have to assign to this node in addition an *information set*. This is a subset of the nodes that are assigned

to i and always includes n itself. It represents the information available to i in situation n. If n' is an element of the information set assigned to n, then the same information set has to be assigned to n'. The idea is: if n and n' are elements of the same information set, then player i cannot distinguish between the two situations, i.e. he does not know whether he is in n_1 or in n_2.

4 There is a set of *outcomes*. To each end node we have to assign exactly one outcome. It is the final state resulting from playing the branch starting from the root node and leading to the end node.

5 For each player i in N there exists a payoff function u_i that assigns a real value to each of the outcomes.

In addition it is assumed that nature, player 0, chooses its moves with certain probabilities:

6 For each node assigned to 0 there is a probability distribution P over its possible moves at this node.

Figure 1.5 shows an example of a game in extensive form with nature as a player. Assume that 1 and 2 want to meet at 5 pm. 1 calls 2 and leaves a message (φ) for her: *"I have an appointment with the doctor. You can pick me up there."* Now assume that 1 has regularly appointments with two different doctors. Where should 2 pick him up? Of course, if 2 knows that the probability ρ for 1 being at s_1 is higher than the probability ρ' for being at s_2, then she should better go to s_1 (g_1) than to s_2 (g_2).

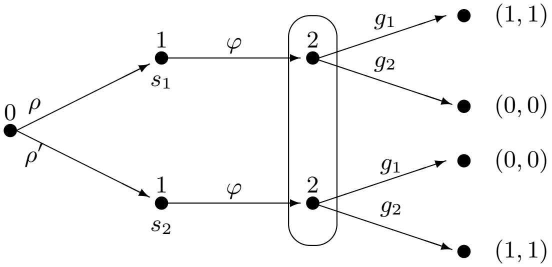


Figure 1.5

In this introduction we concentrated on classical Game Theory . It rests on very strong assumptions about the players' rationality and computational abilities. A Nash equilibrium, to be a really clear cut solution for a game, presupposes that each other's strategies are commonly known. In general, it's not only each other's strategies but also each other's reasoning that must be commonly known. Hence, there have been efforts to develop criteria for players with *bounded rationality*. An extreme reinterpretation is provided by *evolutionary* game theory, where in fact no reasoning is assumed on the side of the players but a form of (evolutionary) learning. This will be the topic of the third section of this introduction. But before we come to that, we will first describe how the tools of standard decision and game theory can be used to study some pragmatic aspects of communication.

2 Communication in games

2.1 Cheap talk games

2.1.1 A sequential formulation

In his classic work on conventions, Lewis (1969) proposed to study communication by means of so-called signaling games, games that are of immediate relevance for linguistics. Meanwhile, extensions of Lewisean signaling games have become very important in economics and (theoretical) biology. In this section we will only consider the simple kind of signaling games that Lewis introduced, games that are now known as cheap talk games. Cheap talk games are signaling games where the messages are not directly payoff relevant. A signaling game with payoff irrelevant messages is a sequential game with two players involved, player 1, the speaker, and player 2, the hearer. Both players are in a particular state, an element of some set T. Player 1 can observe the true state, but player 2 can not. The latter has, however, beliefs about what the true state is, and it is common knowledge between the players that this belief is represented by probability function P_H over T. Then, player 1 observes the true state t and chooses a message m from some set M. After player 2 observes m (but not t), he chooses some action a from a set $\mathcal{A}$, which ends the game. In this sequential game, the hearer has to choose a (pure) strategy that says what she should do as a response to each message, thus a function from M to $\mathcal{A}$, i.e., an element in $[M \to \mathcal{A}]$. Although strictly speaking not necessary, we can represent also the speaker's strategy already as a function, one in $[T \to M]$. In simple communication games, we call these functions the hearer and speaker strategy, respectively, i.e., H and S. The utilities of both players are given by $u_S(t, a)$ and $u_H(t, a)$.

In cheap talk games, the messages are not directly payoff relevant: the utility functions do not mention the messages being used. Thus, the only effect that a message can have in these games is through its information content: by changing the hearer's belief about the situation the speaker (and hearer) is in. If a message can change the hearer's beliefs about the actual situation, it might also change her optimal action, and thus indirectly affect both players' payoffs.

Let us look at a very simple situation where $T = \{t_1, t_2\}$, $M = \{m_1, m_2\}$, and where f is a function that assigns to each state the unique action in $\mathcal{A}$ that is the desired action for both agents. Let us assume the following utility functions (corresponding closely with Lewis's intentions): $u_S(t_i, H_k(m_j)) = 1 = u_H(t_i, H_k(m_j))$, if $H_k(m_j) = f(t_i)$, 0 otherwise. Let us assume that $f(t_1) = a_1$ and $f(t_2) = a_2$, which means that a_1 is the best action to perform in t_1, and a_2 in t_2. The speaker and hearer strategies will be defined as follows:

$$\begin{array}{ll} S_1 = \{\langle t_1, m_1\rangle, \langle t_2, m_2\rangle\} & S_2 = \{\langle t_1, m_2\rangle, \langle t_2, m_1\rangle\} \\ S_3 = \{\langle t_1, m_1\rangle, \langle t_2, m_1\rangle\} & S_4 = \{\langle t_1, m_2\rangle, \langle t_2, m_2\rangle\} \end{array}$$

$$\begin{array}{ll} H_1 = \{\langle m_1, a_1\rangle, \langle m_2, a_2\rangle\} & H_2 = \{\langle m_1, a_2\rangle, \langle m_2, a_1\rangle\} \\ H_3 = \{\langle m_1, a_1\rangle, \langle m_2, a_1\rangle\} & H_4 = \{\langle m_1, a_2\rangle, \langle m_2, a_2\rangle\} \end{array}$$

In the following description, we represent not the actual payoffs in the payoff table, but rather what the agents who are in a particular situation *expect* they will receive. Thus we will give a table for each situation with the payoffs for each agent $e \in \{S, H\}$ determined by

$$U^*_e(t, S_i, H_j) = \sum_{t' \in T} \mu_e(t'|(S_i(t))) \times U_e(t', H_j(S_i(t')))$$

where $\mu_e(t'|S_i(t))$ is defined by means of conditionalization in terms of strategy S_i and the agent's prior probability function P_e as follows:

$$\mu_e(t'|S_i(t)) = P_e(t'|S_i^{-1}(S_i(t)))$$

and where $S_i^{-1}(m)$ is the set of states in which a speaker who uses strategy S_i uses m. Because the speaker knows in which situation he is, for him it is the same as the actual payoff: $U^*_S(t, S_i, H_j) = u_S(t, H_j(S_i(t)))$. For the hearer, however, it is not the same, because if the speaker uses strategy S_3 or S_4 she still doesn't know what the actual state is. Let us assume for concreteness that $P_H(t_1) = \frac{1}{3}$, and thus that $P_H(t_2) = \frac{2}{3}$. This then gives rise to the tables in 1.9 on the next page:

We have boxed the equilibria of the games played in the different situations. We say that strategy combination $\langle S, H\rangle$ is a Nash equilibrium of the whole game iff $\langle S, H\rangle$ is a Nash equilibrium in both situations (see Definition 3). Thus, we see that we have four equilibria: $\langle S_1, H_1\rangle$, $\langle S_2, H_2\rangle$,

Table 1.9: Cheap talk game: asymmetric information

t_1	H_1	H_2	H_3	H_4
S_1	$\boxed{(1\,;\,1)}$	$(0\,;\,0)$	$\boxed{(1\,;\,1)}$	$(0\,;\,0)$
S_2	$(0\,;\,0)$	$\boxed{(1\,;\,1)}$	$\boxed{(1\,;\,1)}$	$(0\,;\,0)$
S_3	$(1\,;\,\frac{1}{3})$	$(0\,;\,\frac{2}{3})$	$(1\,;\,\frac{1}{3})$	$\boxed{(0\,;\,\frac{2}{3})}$
S_4	$(0\,;\,\frac{2}{3})$	$(1\,;\,\frac{1}{3})$	$(1\,;\,\frac{1}{3})$	$\boxed{(0\,;\,\frac{2}{3})}$

t_2	H_1	H_2	H_3	H_4
S_1	$\boxed{(1\,;\,1)}$	$(0\,;\,0)$	$(0\,;\,0)$	$\boxed{(1\,;\,1)}$
S_2	$(0\,;\,0)$	$\boxed{(1\,;\,1)}$	$(0\,;\,0)$	$\boxed{(1\,;\,1)}$
S_3	$(0\,;\,\frac{1}{3})$	$(1\,;\,\frac{2}{3})$	$(0\,;\,\frac{1}{3})$	$\boxed{(1\,;\,\frac{2}{3})}$
S_4	$(1\,;\,\frac{2}{3})$	$(0\,;\,\frac{1}{3})$	$(0\,;\,\frac{1}{3})$	$\boxed{(1\,;\,\frac{2}{3})}$

$\langle S_3, H_4\rangle$ and $\langle S_4, H_4\rangle$. These equilibria can be pictured as in Figure 1.6 on the following page.

2.1.2 A strategic reformulation and a warning

Above we have analyzed the game as a sequential one where the hearer didn't know in which situation she was. Technically, we have analyzed the game as a sequential game with asymmetric information: the speaker knows more than the hearer. However, it is also possible to analyze it as a standard strategic game in which this asymmetry of information no longer plays a role.[3] If we analyze the game in this way, however, we have to change two things: (i) we don't give separate payoff tables anymore for the different situations; and, as a consequence, (ii) we don't look at (expected) utilities anymore of strategy combinations in a particular state, but have to look at expectations with respect to a *common prior* probability function ρ. Before we do this for signaling games, let us first show how a more simple

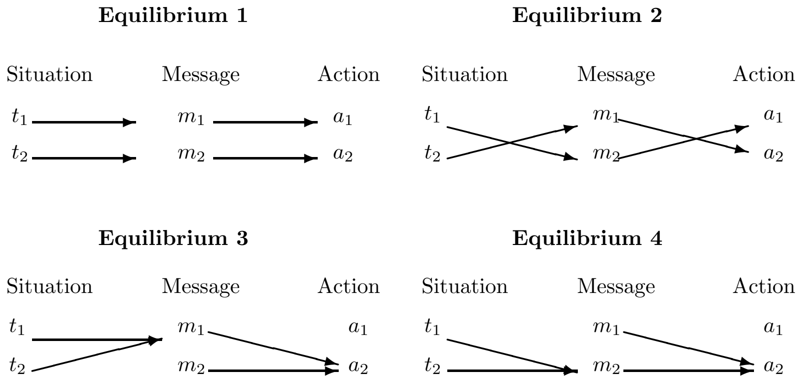


Figure 1.6

game of asymmetric information can be turned into a strategic one with symmetric information.

Suppose that row player knows which situation he is in, but Column-player does not. The reason might be that Column doesn't know what the preferences of Row are. Column thinks that the preferences are as in t_1 or as in t_2. Notice that in both situations the preferences of Column are the same, which represents the fact that Row knows what the preferences are of Column.

Table 1.10: Simple game of asymmetric information

t_1	c_1	c_2
r_1	$(2\,;\,1)$	$(3\,;\,0)$
r_2	$(0\,;\,-1)$	$(2\,;\,0)$

t_2	c_1	c_2
r_1	$(2\,;\,1)$	$(3\,;\,0)$
r_2	$(3\,;\,-1)$	$(5\,;\,0)$

Notice that in t_1, r_1 is the dominant action to perform for Row, while in t_2 it is r_2. Column has to make her choice depending on what she thinks is best. She knows that Row will play r_1 in t_1, and r_2 in t_2 (assuming that Row is rational), but she doesn't know what is the actual situation. However, she has some beliefs about this. Let us assume that $P_{Col}(t_1)$ is the probability with which Column thinks that t_1 is the actual situation (and thus that $P_{Col}(t_2) = 1 - P_{Col}(t_1)$). Then Column will try to maximize her expected utility. Thus, she will choose c_1 in case $EU_{Col}(c_1) > EU_{Col}(c_2)$, and c_2 otherwise (in case $EU_{Col}(c_1) = EU_{Col}(c_2)$ she doesn't care.) Notice that $EU_{Col}(c_1) > EU_{Col}(c_2)$ if an only if $[(P_{Col}(t_1) \times 1) + (P_{Col}(t_2) \times (-1))] > [(P_{Col}(t_1) \times 0) + (P_{Col}(t_2) \times 0)]$ if and only if $P_{Col}(t_1) > P_{Col}(t_2)$. Thus, we predict by using the Nash equilibrium solution concept that Row will play r_1 in situation t_1, and r_2 in situation t_2, and Column will play c_1 if $P_{Col}(t_1) \geq \frac{1}{2}$ and c_2 if $P_{Col}(t_1) \leq \frac{1}{2}$. We can characterize the Nash equilibria of the game also as follows: $\langle(r_1, r_2), c_1\rangle$ iff $P_{Col}(t_1) \geq \frac{1}{2}$ and $\langle(r_1, r_2), c_2\rangle$ iff $P_{Col}(t_1) \leq \frac{1}{2}$, where $\langle(r_1, r_2), c_1\rangle$, for instance, means that Row will play r_1 in situation t_1 and r_2 in situation t_2.

This game of asymmetric information can also be studied as a standard strategic game of symmetric, or complete, information, if we assume that the participants of the game can represent their beliefs in terms of a *common prior* probability function, ρ, and that Row has obtained his complete information via updating this prior probability function with some extra information by means of standard conditionalization (Harsanyi 1967-1968). In our case we have to assume that the common prior probability function ρ is just the same as Column's probability function P, and that Row has received some information X (in state t_1, for instance) such that after conditionalization ρ with this information, the new probability function assigns the value 1 to t_1, i.e. $\rho(t_1/X) = 1$.

If we analyze our example with respect to a common prior probability function, the Row-player also has to pick his strategy *before* he finds out in which situation the game is being played, and we also have to define his payoff function with respect to the prior probability function $\rho = P_{Col}$. But this means that the actions, or *strategies*, of Row player are now *functions* that tell him what to do in each situation. In the game described above Row now has to choose between the following *four* strategies:

$r_{11} = r_1$ in t_1 and r_1 in t_2 $\qquad$ $r_{22} = r_2$ in t_1 and r_2 in t_2
$r_{12} = r_1$ in t_1 and r_2 in t_2 $\qquad$ $r_{21} = r_2$ in t_1 and r_1 in t_2

This gives rise to the strategic game from Table 1.11 on the next page with respect to the common prior probability function ρ that assigns the value $\frac{2}{3}$ to t_1.

Now we see that $\langle r_{12}, c_1\rangle$ is the Nash equilibrium. In a game of asymmetric information, this equilibrium corresponds to equilibrium $\langle(r_1, r_2), c_1\rangle$, which is exactly the one we found above if $P_{Col}(t_1) = \frac{2}{3}$. So, we have seen that our game of asymmetric information gives rise to the same equilibrium as its reformulation as a strategic game with symmetric information. This property, in fact, is a general one, and not limited to this simple example,

Table 1.11: Strategic reformulation of simple game

	c_1	c_2
r_{11}	$(2\,;\,1)$	$(3\,;\,0)$
r_{22}	$(1\frac{1}{2}\,;\,-1)$	$(3\frac{1}{2}\,;\,0)$
r_{12}	$\boxed{(2\frac{1}{2}\,;\,\frac{1}{3})}$	$(4\,;\,0)$
r_{21}	$(1\,;\,-\frac{1}{3})$	$(2\frac{1}{2}\,;\,0)$

but also extends to sequential games.

In section 2.1.1 we have introduced signaling games as (simple) sequential games where the speaker has some information that the hearer does not have. Exactly as in the just described simple example, however, we can turn this game into a strategic one with symmetric information if we assume a *common prior* probability function ρ, such that $\rho = P_H$. The payoffs are now determined as follows: $\forall a \in \{S, H\} : U^*_a(S_i, H_j) = \sum_{t \in T} \rho(t) \times u_a(t, H_j(S_i(t)))$. As a result, the game analyzed as one of symmetric information receives the following completely symmetric strategic payoff table:

Table 1.12: Signaling game: strategic reformulation

	H_1	H_2	H_3	H_4
S_1	$\boxed{(1\,;\,1)}$	$(0\,;\,0)$	$(\frac{1}{3}\,;\,\frac{1}{3})$	$(\frac{2}{3}\,;\,\frac{2}{3})$
S_2	$(0\,;\,0)$	$\boxed{(1\,;\,1)}$	$(\frac{1}{3}\,;\,\frac{1}{3})$	$(\frac{2}{3}\,;\,\frac{2}{3})$
S_3	$(\frac{1}{3}\,;\,\frac{1}{3})$	$(\frac{2}{3}\,;\,\frac{2}{3})$	$(\frac{1}{3}\,;\,\frac{1}{3})$	$(\frac{2}{3}\,;\,\frac{2}{3})$
S_4	$(\frac{2}{3}\,;\,\frac{2}{3})$	$(\frac{1}{3}\,;\,\frac{1}{3})$	$(\frac{1}{3}\,;\,\frac{1}{3})$	$(\frac{2}{3}\,;\,\frac{2}{3})$

Strategy combination $\langle S, H\rangle$ is a Nash equilibrium (cf. Definition 3) in this kind of game given probability function ρ if (i) $\forall S' : U^*_S(S, H) \geq U^*_S(S', H)$ and (ii) $\forall H' : U^*_H(S, H) \geq U^*_H(S, H')$. We see that this game has four equilibria, $\langle S_1, H_1\rangle$, $\langle S_2, H_2\rangle$, $\langle S_3, H_4\rangle$, $\langle S_4, H_4\rangle$, which again exactly correspond to the earlier equilibria in the game of *a*symmetric information.

We mentioned already that Lewis (1969) introduced cheap talk games to explain the use and stability of linguistic conventions. More recently, these, or similar, type of games have been used by, among others, Prashant Parikh (Parikh 1991, Parikh 2001), de Jaegher (de Jaegher 2003) and van Rooij (van Rooij 2004) to study some more concrete pragmatic phenomena of language use.[4]

Until now we have assumed that all Nash equilibria of a cheap talk game are the kind of equilibria we are looking for. However, in doing so we missed a lot of work in economics discussing refinements of equilibria. In order to look at games from a more fine-grained point of view, we have to look at the game as a sequential one again.

Remember that when we analyzed cheap talk games from a sequential point of view, we said that $\langle S_i, H_j \rangle$ is a Nash equilibrium with respect to probability functions P_a if for each t, S_k and H_m:

$$U_S^*(t, S_i, H_j) \geq U_S^*(t, S_k, H_j)$$

and

$$U_H^*(t, S_i, H_j) \geq U_H^*(t, S_i, H_m)$$

For both agents a, we defined $U_a^*(t, S_i, H_j)$ as follows: $\sum_{t' \in T} \mu_a(t'|S_i(t)) \times U_a(t', H_j(S_i(t')))$, where $\mu_a(t'|(S_i(t)))$ was *defined* in terms of the agents' prior probability functions P_a and standard conditionalization as follows: $\mu_a(t'|(S_i(t))) = P_a(t'|S^{-1}(S(t)))$. On this analysis, the equilibrium is the same as the one used in the strategic game.

However, assuming a conditional probability function μ to be defined in this way gives in many games rise to counterintuitive equilibria. In the context of cheap talk games, all of them have to do with how the hearer would react to a message that is not sent in the equilibrium play of the game, i.e., not sent when the speaker uses strategy S_i that is part of the equilibrium. Notice that if m is such a message, $\mu_H(t|m)$ can take *any* value in $[0, 1]$ according to the above definition of the equilibrium, because the conditional probability function is not defined then.

To give a very simple (though, admittedly, a somewhat unnatural) illustration of the kind of problems that might arise, suppose we have a signaling game with $T = \{t_1, t_2\}$, and $M = \{m, \varepsilon\}$ as in Table 1.13 on the following page. Let us also assume that m is a message with a fully underspecified meaning, but that, for some reason, the speaker could use ε only in situation t_2, although the hearer still might react in any possible way. Then we have the following strategies.

Now let us again assume for all $a \in \{S, H\}$ that $U_a(t, H(S(t))) = 1$ if $H(S(t)) = f(t)$, 0 otherwise, with f defined as in section 2.1.1. Now

Table 1.13: Game where standard conditionalization is not good enough

	t_1	t_2
S_1	m	ε
S_2	m	m

	m	ε
H_1	a_1	a_1
H_2	a_1	a_2
H_3	a_2	a_1
H_4	a_2	a_2

we see that if $P_H(t_2) > \frac{1}{2}$ as before, we have the following equilibria: $\langle S_1, H_2\rangle$, $\langle S_2, H_3\rangle$, and $\langle S_2, H_4\rangle$. The first one is independent of the probability function, while the latter two hold if we assume that $\mu_H(t_1|m) \geq \frac{1}{2}$ and $\mu_H(t_2|m) \geq \frac{1}{2}$, respectively. The equilibrium $\langle S_2, H_4\rangle$ seems natural, because ε is interpreted by the receiver as it might be used by the sender. Equilibrium $\langle S_2, H_3\rangle$, however, is completely unnatural, mainly because if $P_H(t_2) > \frac{1}{2}$, it seems quite unnatural to assume that $\mu_H(t_1|m) \geq \frac{1}{2}$ is the case. To account for much more complicated examples, Kreps and Wilson (1982) propose that instead of *defining* $\mu_H(t|m)$ in terms of a speaker strategy S and prior probability function P_H, the conditionalization requirement is only a *condition* on what $\mu_H(t|m)$ should be, in case m is used according to the speaker's strategy. This leaves room for an extra constraint on what $\mu_H(t|m)$ should be in case m is not used by the speaker's strategy. The extra condition Kreps and Wilson propose is their *consistency condition* for beliefs at information sets that are not reached in equilibrium. It says, roughly speaking, that for each situation t_i and message m_i that is not sent according to the speaker's strategy, the posterior probability of t_i at the information state that results after m_i would be sent, i.e. $\mu_H(t_i|m_i)$, should (in simple cases) be as close as possible to the prior probability of t_i. In our case this means that, although ε is not sent if the speaker uses strategy S_2, we should still give $\mu_H(t_1|m)$ and $\mu_H(t_2|m)$ particular values, namely $\mu_H(t_1|m) = P_H(t_1) < \frac{1}{2}$ and $\mu_H(t_2|m) = P_H(t_2) > \frac{1}{2}$. But if the beliefs are like this, the Nash equilibrium $\langle S_2, H_3\rangle$ ceases to be an equilibrium anymore when taking this consistency requirement into account, and we have explained why it is unnatural.

In this introduction we won't bother anymore with the above mentioned refinements, so we might as well think of the game from a strategic point of view.

2.2 The quantity of information transmission

Notice that in the first two equilibria of the cheap talk game described in section 2.1.1. there is a 1-1-1 correspondence between situations, messages and actions, and it is natural to say that if speaker and hearer coordinate on the first equilibrium, the speaker uses message m_i to indicate that she is in situation t_i and wants the hearer to perform action a_i. As a natural special case we can think of the actions as ones that interpret the messages. In these cases we can identify the set of actions, $\mathcal{A}$ (of the hearer), with the set of states, T.[5] In that case it is most natural to think of the set of states as the set of *meanings* that the speaker wants to express, and that in the first two equilibria there exists a 1-1 correspondence between meanings and messages. One of the central insights of Lewis' (1969) work on conventions was that the meaning of a message can be defined in terms of the game theoretical notion of an equilibrium in a signaling game: messages that don't have a pre-existing meaning acquire such a meaning through the equilibrium play of the game. Lewis (1969) proposes (at least when ignoring context-dependence) that all and only all equilibria where there exists such a 1-1 correspondence, which he calls *signaling systems*, are appropriate candidates for being a *conventional* solution for communicating information in a signaling game.

In contrast to the first two equilibria, no information transmission is going on in equilibria 3 and 4. Still, they count as equilibria: in both equilibria the hearer is justified in ignoring the message being sent and always plays the action which has the highest expected utility, i.e. action a_2, if the speaker sends the same message in every situation; and the speaker has no incentive to send different messages in different states if the hearer ignores the message and always performs the same action. Thus, we see that in different equilibria of a cheap talk communication game, different amounts of information can be transmitted. But for cheap talk to allow for informative communication at all, a speaker must have different preferences over the hearer's actions when he is in different states. Likewise, the hearer must prefer different actions depending on what the actual situation is (talk is useless if the hearer's preferences over actions are independent of what the actual situation is.) Finally, the hearer's preferences over actions must not be completely opposed to that of the speaker's. These three conditions are obviously guaranteed if we assume, as in the example above, that there is perfect alignment of preferences between speaker and hearer, and for both speaker and hearer there is a one-to-one relation between states and optimal actions (in those states). In general, however, we don't want such an idealistic assumption. This gives rise to the question how informative cheap talk can be. That is, how fine-grained can and will the speaker reveal the true situation if talk is cheap? We will address this issue in a moment, but first would like to say something about a refinement of the Nash equilibrium concept in sequential games that sometimes helps us to eliminate some counterintuitive equilibria.

As mentioned above, the main question asked in cheap talk games is how much information can be transmitted, given the preferences of the different agents. In an important article, Crawford and Sobel (1982) show that the amount of credible communication in these games depends on how far the preferences of the participants are aligned. To illustrate, assume that the state, message and action spaces are continuous and between the interval of zero and one. Thus, $T = [0,1]$; the message space is the type space ($M = T$), i.e., $M = [0,1]$, and also the action space is in the interval $[0,1]$. Now, following Gibson (1992), we can construct as a special case of their model the following quadratic utility functions for speaker and hearer such that there is a single parameter, $b > 0$, that measures how closely the preferences of the two players are aligned:

$$\begin{array}{rcl} U_H(t,a) & = & -(a-t)2 \\ U_S(t,a) & = & -[a-(t+b)]^2 \end{array}$$

Now, when the actual situation is t, the hearer's optimal action is $a = t$, but the speaker's optimal action is $a = t + b$. Thus, in different situations the speaker has different preferences over the hearer's actions (in 'higher' situations speakers prefer higher actions), and the interests of the players are more aligned in case b comes closer to 0. Crawford and Sobel (1982) show that in such games all equilibria are *partition equilibria* ; i.e., the set of situations T can be partitioned into a finite number of intervals such that senders in a state belonging to the same interval send a common message and receive the same action. Moreover, they show that the amount of information revealed in equilibrium increases as the preferences of the speaker and the hearer are more aligned. That is, if parameter b approaches 0, there exists an equilibrium where the speaker will tell more precisely which situation he is in, and thus more communication is possible. However, when parameter b has the value 1, it represents the fact that the preferences of speaker and hearer are opposed. A speaker in situation $t = 1$, for instance prefers most action $a = 1$, and mostly disprefers action $a = 0$. If $b = 1$, however, a hearer will prefer most action $a = 0$ and most dislikes action $a = 1$. As a result, no true information exchange will take place if $b = 1$, i.e., if the preferences are completely opposed.

To establish the fact proved by Crawford and Sobel, no mention was made of any externally given meaning associated with the messages. What happens if we assume that these messages in fact *do* have an externally given meaning, taken to be sets of situations? Thus, what happens when we adopt an externally given interpretation function $[\cdot]$ that assigns to every $m \in M$ a subset of T? The interesting question is now not whether the game has equilibria in which we can associate meanings with the messages, but rather whether there exist equilibria where the messages are sent in a *credible* way. That is, are there equilibria where a speaker sends a message with meaning $\{t_i\}$ if and only if she is in state t_i? As it turns out, the old question concerning informative equilibria in signaling games without pre-existing meaning and the new one concerning credible equilibria in signaling games with messages with pre-existing meaning are closely related. Consider a two-situation two-action game with the following utility table. ("t_H" and "t_L" are mnemonic for "high type" and "low type", which mirror the preferences of the receiver.)

Table 1.14: Two-situation, two action

	a_H	a_L
t_H	$(1\,;\,1)$	$(0\,;\,0)$
t_L	$(1\,;\,0)$	$(0\,;\,1)$

In this game, the informed sender prefers, irrespective of the situation he is in, column player to choose a_H, while column player wants to play a_H if and only if the sender is in situation t_H. Now assume that the expected utility for the hearer to perform a_H is higher than that of a_L (because $P(t_H) > P(t_L)$). In that case, in both situations speakers have an incentive to send the message that conventionally expresses $\{t_H\}$. But this means that in this game a speaker in t_L has an incentive to lie, and thus that the hearer cannot take the message to be a credible indication that the speaker is in situation t_H, even if the speaker was actually in that situation. Farrell (1988, 1993) and Rabin (1990) discussed conditions under which messages with a pre-existing meaning can be used to credibly transmit information. They show that this is possible by requiring that the hearer believes what the speaker says if it is in the latter's interest to speak the truth. The paper of Stalnaker in this volume explores the connection between this work on credible information transmission and Gricean pragmatics.

We have indicated above that the assumption that messages have an externally given pre-existing meaning doesn't have much effect on the equilibria of cheap talk games, or on Crawford and Sobel's (1982) result on the amount of possible communication in such games. This holds at least if no requirements are made on what should be believed by the hearer, and if no constraints are imposed on what kinds of meanings can be expressed in particular situations, for instance if no requirements like $\forall t \in T : t \in [S(t)]$, saying that speakers have to tell the truth, are put upon speakers' strategies S.

2.3 Verifiable communication with a skeptic audience

What happens if we *do* put extra constraints upon what can and what cannot be said? As it turns out, this opens up many new possibilities of credible communication. In fact, Lipman and Seppi (1995) (summarized in Lipman 2003) have shown that with such extra constraints, interesting forms of reliable information transmission can be predicted in games where you expect it the least: in debates between agents with opposing preferences.

Before we look at debates, however, let us first consider cheap talk games when we assume that the signals used come with a pre-existing meaning and, moreover, that speakers always tell the truth. This still doesn't guarantee that language cannot be used to mislead one's audience. Consider again the two-situation two-action game described above, but now assume in addition that we demand that the speaker speaks the truth: $t_i \in [S(t_i)]$. The rational message for an individual in the 'high' situation to send is still one that conventionally expresses $\{t_H\}$, but an individual in the 'low' situation now has an incentive to send a message with meaning $\{t_H, t_L\}$. If the hearer is naive she will choose a_H after hearing the signal that expresses $\{t_H, t_L\}$, because a_H has the highest expected utility. A more sceptical hearer, however, will argue that a speaker that sends a message with meaning $\{t_H, t_L\}$ must be one that is in a 'low' situation, because otherwise the speaker could, and thus should (in her own best interest) have sent a message with meaning $\{t_H\}$. Thus, this sceptical hearer will reason that the speaker was in fact in a low-type situation and interprets the message as $\{t_L\}$. Indeed, this game has an equilibrium where the speaker and hearer act as described above. In general, suppose that the speaker has the following preference relation over a set of 10 situations: $t_1 < t_2 < ... < t_{10}$ (meaning that t_1 is the worst situation) and sends a message m with pre-existing meaning $[m]$. A sceptical hearer would then assign to m the following pragmatic interpretation $S(m)$ based on the speaker's preference relation '$<$', on the assumption that the speaker knows which situation he is in:

$$S(m) \quad = \quad \{t \in [m] | \neg\exists t' \in [m] : t' < t\}$$

This pragmatic interpretation rule is based on the assumption that the speaker gives as much information as he can that is useful to him, and that the hearer anticipates this speaker's maxim (to be only unspecific with respect to more desirable states) by being sceptical when the speaker gives a message with a relatively uninformative meaning.[6]

Now consider debates in which the preferences of the participants are mutually opposed. Suppose that debaters 1 and 2 are two such players who *both* know the true state. Now, however, there is also a third person, the *observer*, who doesn't. Both debaters present evidence to the observer, who then chooses an action $a \in A$ which affects the payoffs of all of them. We assume that the observer's optimal action depends on the state, but that the preferences of the debaters do not. In fact, we assume that the preferences of debaters 1 and 2 are strictly opposed: in particular, if debater 1 prefers state t_i above all others, t_i is also the state that debater 2 dislikes most. By assuming that the utility functions of all three participants are of type $U_j(t, a)$, we again assume that the message being used is not directly payoff relevant, just as in cheap talk games.

We assume that each debater can send a message. Let us assume that S denotes the strategy of debater 1, and R the strategy of debater 2. In contrast to the cheap talk games discussed above, we now crucially assume that the messages have an externally given meaning given by interpretation function $[\cdot]$. Let us first assume that while debater 1 can make very precise statements, i.e., that a particular state t holds, debater 2 can only make very uninformative statements saying that a particular state is *not* the case. Let us assume for concreteness that $T = \{t_1, ..., t_{10}\}$. Then the 'meaning' of $S(t_i)$, $[S(t_i)]$ can consist of one state, $\{t_j\}$, while the meaning of $R(t_i)$, $[R(t_i)]$, always consists of 9 states. Thus, debater 1 can be much more informative about the true state, and is thus in the advantage. But debater 2 has an advantage over debater 1 as well: in contrast to what is known about debater 1, it is commonly known of debater 2 that she is reliable and will only make *true* statements. Thus, for all $t_i \in T : t_i \in [R(t_i)]$, while it might be that $t_i \notin [S(t_i)]$.

Suppose now that the observer may ask two statements of the players. The question is, how much information can the observer acquire? One is tempted to think that the messages cannot really give a lot of information: debater 1 has no incentive to tell the truth, so acquiring two messages from him is completely uninformative. Debater 2 will provide true information, but the informative value of her messages is very low: after two messages from her the observer still doesn't know which of the remaining 8 states is the true one. Surprisingly enough, however, Lipman and Seppi (1995) show that the observer can organize the debate such that after two rounds of communication, he knows for certain which state actually obtains.

The trick is the following: the observer first promises, or warns, debater 1 that in case he finds out that the latter will not give a truthful message, he will punish debater 1 by choosing the action that is worst for him. This is possible because it is common knowledge what the agents prefer. For concreteness, assume that debater 1 has the following preferences $t_{10} > t_9 > ... > t_1$. Afterwards, the observer first asks debater 1 which state holds, and then asks debater 2 to make a statement. Suppose that the first debater makes a very informative statement of the form 'State t_i is the true state'. Obviously, debater 2 will refute this claim if it is false. For in that case the observer will as a result choose the state most unfavorable to debater 1, and thus most favorable to debater 2, i.e. t_1. Thus, if he is precise, debater 1 has an incentive to tell the true state, and the observer will thus learn exactly which state is the true one. Suppose that the true state is the one most undesirable for debater 1, t_1. So, or so it seems, he has every reason to be vague. Assume that debater 1 makes a vague statement with meaning $\{t_i, ..., t_n\}$. But being vague now doesn't help: if the true state is ruled out by this vague meaning, debater 2 will claim that (even) the least preferred state in it is not true, and if debater 2 doesn't refute debater 1's claim in this way the observer will choose the most unfavorable state for debater 1 compatible with the true message with meaning $\{t_i, ..., t_n\}$. In general, if debater 1's message m has meaning $[m]$, and if m is not refuted, then the observer will 'pragmatically' interpret m as follows: $\{t \in [m] | \neg\exists t' \in [m] : t' < t\}$, where '$t' < t$' means that debater 1 (strictly) prefers t to t'. Notice that this is exactly the pragmatic interpretation rule $S(m)$ described above. From a signaling game perspective, this just means that the games has a completely separating equilibrium: whatever the true state is, it is never in the interest of debater 1 not to say that this is indeed the true state.

The example discussed above is but a simple, special case of circumstances characterized by Lipman and Seppi (1995) in which observers can 'force' debaters to provide precise and adequate information, even though they have mutually conflicting preferences.

The discussion in this section shows that truthful information transmission is possible in situations in which the preferences of the conversational participants are mutually opposed. This seems to be in direct conflict with the conclusion reached in section 2.2, where it was stated that credible information transmission is impossible in such circumstances. However, this conflict is not real: on the one hand, a central assumption in cheap talk games is that talk is really cheap: one can say what one wants because the messages are *not verifiable*. The possibility of credible information transmission in debates, on the other hand, crucially depends on the assumption that claims of speakers are verifiable to at least some extent, in other words, they are falsifiable (by the second debater), and that outside observers can punish the making of misleading statements. In fact, by taking the possibility of falsification and punishment into account as well, we predict truthful communication also in debates, because the preferences of the agents which seemed to be opposing are still very much aligned at a 'deeper' level.

This subsection also shows that if a hearer knows the preferences of the speaker and takes him to be well-informed, there exists a natural 'pragmatic' way to interpret the speaker's message which has already a pre-existing 'semantic' meaning, based on the assumption that speakers are rational and only unspecific, or vague, with respect to situations that are more desirable for them.

2.4 Debates and Pragmatics

It is well established that a speaker in a typical conversational situation communicates more by the use of a sentence than just its conventional truth conditional meaning. Truth conditional meaning is enriched with what is *conversationally implicated* by the use of a sentence. In pragmatics – the study of language use – it is standard to assume that this way of enriching conventional meaning is possible because we assume speakers to conform to Grice's (1967) *cooperative principle*, the principle that assumes speakers to be rational cooperative language users. This view on language use suggests that the paradigmatic discourse situation is one of cooperative information exchange.

Merin (1999b) has recently argued that this view is false, and hypothesized that discourse situations are paradigmatically ones of explicit or tacit debate. He bases this hypothesis on the work of Ducrot (1973) and Anscombre and Ducrot (1983) where it is strongly suggested that some phenomena troublesome for Gricean pragmatics can be analyzed more successfully when we assume language users to have an *argumentative* orientation. In the sequel we will sketch some of Ducrot's arguments for such an alternative view on communication, and we will describe Merin's analysis of some implicatures which are taken to be troublesome for a cooperative view on language use.

Adversary connectives The connective *but* is standardly assumed to have the same truth-conditional meaning as *and*. Obviously, however, they are used differently. This difference is accounted for within pragmatics. It is normally claimed that 'A and B' and 'A but B' give rise to different *conventional implicatures*, or appropriateness conditions. On the basis of sentences like (1) it is normally (e.g. Frege 1918) assumed that sentences of the form 'A but B' are appropriate, if B is unexpected given A.

(1) John is tall but no good at baseball.

This, however, cannot be enough: it cannot explain why the following sentence is odd:

(2) John walks but today I won the jackpot.

Neither can it explain why the following sentence is okay, because expensive restaurants are normally good.

(3) This restaurant is expensive, but good.

Ducrot (1973), Anscombre and Ducrot (1983) and Merin (1999a,b) argue that sentences of the form 'A but B' are always used argumentatively, where A and B are arguments for complementary conclusions: they are contrastive in a rhetorical sense. For instance, the first and second conjunct of (3) argue in favor of not going and going to the restaurant, respectively.

Not only sentences with *but*, but also other constructions can be used to express a relation of rhetorical contrast (cf. Horn 1991). These include complex sentences with *while*, *even if*, or *may* in the first clause (i.e. concession), and/or *still*, *at least*, or *nonetheless* either in place of or in addition to the *but* of the second clause (i.e. affirmation):

(4) a. While she won by a {small, *large} margin, she did win.

b. Even if I have only three friends, at least I have three.

c. He may be a professor, he is still an idiot.

Anscombre and Ducrot (1983) argue that rhetorical contrast is not all to the appropriateness of sentences like (3) or (4a)-(4c). It should also be the case that the second conjunct should be an argument in favor of conclusion H that the speaker wants to argue for. And, if possible, it should be a stronger argument for H than the first conjunct is for $\overline{H}$. In terms of Merin's notion of relevance discussed in section 1.1 this means that the conjunction 'A but B' is appropriate only if $r_H(A) < 0$, $r_H(B) > 0$ and $r_H(A \wedge B) > 0$. In this way it can also be explained why (3), for instance, can naturally be followed by H = 'You should go to that restaurant', while this is not a good continuation of

(5) This restaurant is good, but expensive.

which is most naturally followed by $\overline{H}$ = 'You should not go to that restaurant'.

There is another problem for a standard Gricean approach to connectives like *but* that can be solved by taking an argumentative perspective (cf. Horn 1991). It seems a natural rule of cooperative conversation not to use a conjunctive sentence where the second conjunct is entailed by the first. Thus, (6) is inappropriate:

(6) *She won by a small margin, and win she did.

However, even though *but* is standardly assumed to have the same truth-conditional meaning as *and*, if we substitute *and* in (6) by *but*, the sentence becomes perfectly acceptable:

(7) She won by a small margin, but win she did.

If – as assumed by standard Gricean pragmatics – only truth-conditional meaning is taken as input for pragmatic reasoning, it is not easy to see how this contrast can be accounted for. By adopting Ducrot's hypothesis that in contrast to (6) the conjuncts in (7) have to be rhetorically opposed, the distinction between the two examples can be explained easily: if a speaker is engaged in a debate with somebody who argued that Mrs. X has a relative lack of popular mandate, she can use (7), but not (6).

Merin's (1999a) formalization allows him also to explain in a formal rigorous manner why (7) 'She won by a small margin, but win she did' can be appropriate, although the second conjunct is entailed by the first. The possibility of explaining sentences like (7) depends on the fact that even if B is entailed by A, $A \models B$, it is still very well possible that there are H and probability functions in terms of which $r_{\cdot}(\cdot)$ is defined such that $r_H(A) < r_H(B)$. Thus, the notion of relevance used by Merin does not increase with respect to the entailment relation.

Thus, it appears that an argumentative view on language use can account for certain linguistic facts for which a non-argumentative view seems problematic.

Scalar reasoning Anscombre and Ducrot (1983) and Merin (1999b) argue that to account for so-called 'scalar implicatures' an argumentative view is required as well. Scalar implicatures are normally claimed to be based on Grice's *maxim of quantity*: the requirement to give as much information as is required for the current purposes of the exchange. On its standard implementation, this gives rise to the principle that everything 'higher' on a scale than what is said is false, where the ordering on the scales is defined in terms of informativity. Standardly, scales are taken to be of the form $\langle P(k), ..., P(m)\rangle$, where P is a simple predicate (e.g. *Mary has x children*) and for each $P(i)$ higher on the scale than $P(j)$, the former must be more informative than the latter. From the assertion that $P(j)$ is true we then conclude by scalar implicature that $P(i)$ is false. For instance, if Mary says that she has two children, we (by default) conclude that she doesn't have three children, because otherwise she could and should have said so (if the number of children is under discussion). Other examples are scales like $\langle A \wedge B, A \vee B \rangle$: from the claim that John *or* Mary will come, we are normally allowed to conclude that they will not both come.

Unfortunately, as observed by Fauconnier (1975), Hirschberg (1985) and others, we see inferences from what is not said to what is false very similar to the ones above, but where what is concluded to be false is *not more informative* than, or does not entail, what is actually said. For instance, if Mary answers at her job-interview the question whether she speaks French by saying that her husband does, we conclude that she doesn't speak French herself, although this is not semantically entailed by Mary's answer. Such scalar inferences are, according to Anscombre and Ducrot (1983) best accounted for in terms of an argumentative view on language: Mary wants to have the job, and for that it would be more useful that she herself speaks French than that her husband does. The ordering between propositions should not be defined in terms of informativity, or entailment, but rather in terms of argumentative force. Thus, from Mary's claim that her husband speaks French we conclude that the proposition which has a higher argumentative value, i.e., that Mary speaks French herself, is false. It would be obvious how to account for this in terms of the relevance function used by Merin: assume that H is the proposition 'Mary gets the job'.

Perhaps surprisingly, this natural reasoning schema is *not* adopted in Merin (1999b). In fact, he doesn't want to account for conversational implicatures in terms of the standard principle that everything is false that the speaker didn't say, but could have said. Instead, he proposes to *derive* scalar implicatures from the assumption that conversation is *always* a game in which the preferences of the agents are diametrically opposed. From this view on communication, it follows that assertions and concessions have an 'at least' and 'at most' interpretation, respectively:

> if a proponent, Pro, makes a claim, Pro won't object to the respondent, Con, conceding more, i.e. a windfall to Pro, but will mind getting less. Con, in turn, won't mind giving away less than conceded, but will mind giving away more. Put simply: claims are such as to engender intuitions glossable 'at least'; concessions, dually, 'at most'. (Merin 1999b, p. 191).

This intuition is formalized in terms of Merin's definition of *relevance cones* defined with respect to contexts represented as $\langle P, h \rangle$ (I minimally changed Merin's (1999b) actual definition 8 on page 197.)

Definition 5 The *upward (relevance) cone* $^{\geq S}\phi$ of an element ϕ of a subset $S \subseteq F$ of propositions in context $\langle P, h \rangle$ is the union of propositions in S that are at least as relevant to H with respect to P as ϕ is. The *downward (relevance) cone* $^{\leq S}\phi$ of ϕ in context $\langle P, H \rangle$ is, dually, the union of S-propositions at most as relevant to H with respect to P as ϕ is.

On the basis of his view of communication as a (bargaining) game with opposing preferences, Merin hypothesizes that while the upward cone of a proposition represents Pro's claim, the downward cone represents Con's default expected compatible counterclaim (i.e., *concession*). Net meaning, then is proposed to be the intersection of Pro's claim and Con's counterclaim: ${}^{\geq S}\phi \, \cap \, {}^{\leq S}\phi$, the intersection of what is asserted with what is conversationally implicated.

Now consider the particularized scalar implicature due to Mary's answer at her job interview to the question whether she speaks French by saying that her husband does. As suggested above, the goal proposition, H, now is that Mary gets the job. Naturally, the proposition a = [Mary speaks French] has a higher relevance than the proposition B = [Mary's husband speaks French]. The net meaning of Mary's actual answer is claimed to be ${}^{\geq S}B \cap {}^{\leq S}B$. This gives rise to an appropriate result if we rule out that $B \in S$. This could be done if we assume that S itself *partitions* the state space (in fact, this is what Merin normally assumes). Presumably, this partition is induced by a question like *Who speaks French?* On this assumption it indeed follows that the elements of the partition compatible with A = [Mary speaks French] are not compatible with the downward cone of B, and thus are ruled out correctly.

As we already indicated above, Merin's analysis of conversational implicatures – which assumes that conversation is a bargaining game – is not the only one possible, and perhaps not the most natural one either. In section 2.2 we saw that it makes a lot of sense to assume that (truthful) speakers say as much as they can about situations that are desirable for them. In case the speaker is taken to be well-informed, we can conclude that what speakers do not say about desirable situations is, in fact, not true (if the speaker is taken to be knowledgeable about the relevant facts). We formulated a 'pragmatic' interpretation rule for sceptical hearers that have to 'decode' the message following this reasoning, to hypothesize what kind of situation the speaker is in. Now consider the example again that seemed similar to scalar reasoning but could not be treated in that way in standard Gricean analyses: the case where Mary answers at her job-interview the question of whether she speaks French by saying that her husband does. Intuitively, this gives rise to the scalar implicature that the 'better' answer, that Mary herself speaks French, is false. As already suggested above, this example cannot be treated as a scalar implicature in the standard implementation of Gricean reasoning because the proposition that Mary speaks French is not more informative than, or does not entail, the proposition that her husband does. But notice that if we assume the scale to be the preference order (between states) of the speaker, we can account for this example in terms of our earlier mentioned pragmatic interpretation rule. All we have to assume for this analysis to work is that the state where speaker Mary speaks French herself is more preferred to one where she does not. Thus, we can account for the particularized conversational implicature that Mary doesn't speak French in terms of the pragmatic interpretation rule described in section 2.2.

The pragmatic interpretation rule that we used above is not only relevant for cases where speaker and hearer have opposing preferences (to at least some degree), but is also perfectly applicable in ideal Gricean circumstances where the preferences of the agents are well-aligned.[7] Thus, even if neither the Gricean cooperative view on language use, nor the alternative argumentative view has universal applicability, this doesn't mean that conversational implicatures cannot still be accounted for by means of a general rule of interpretation.

3 Evolutionary game theory

3.1 The evolutionary interpretation of game theory

The classical interpretation of game theory makes very strong idealization about the rationality of the players. First, it is assumed that every player is logically omniscient. The players are assumed to know all logical theorems and all logical consequences of their non-logical beliefs. Second, they are assumed to always act in their enlightened self interest (in the sense of utility maximization). Last but not least, for a concept like "Nash equilibrium" to make sense in classical GT, it has to be common knowledge between the players (a) what the utility matrix is like, and (b) that all players are perfectly rational. Each player has to rely on the rationality of the others without doubt, he has to rely on the other players relying on his own rationality etc. These assumptions are somewhat unrealistic, and variants of classical Game Theory that try to model the behavior of real people in a less idealized way are therefore of high relevance.

A substantial part of current game theoretical research is devoted to *bounded rationality*: versions of GT where the above-mentioned rationality assumptions are weakened. The most radical version of this program is *evolutionary game theory* (EGT). It builds on the fundamental intuition that in games that are played very often (as for instance dialogues), strategies that lead to a high payoff at a point in time are more likely to be played in subsequent games than less successful strategies. No further assumptions are made about the rationality of the agents. Perhaps surprisingly, the solution concepts of classical GT are not invalidated by this interpretation but only slightly refined and modified. Originally EGT was developed by theoretical biologists, especially John Maynard Smith (cf. Maynard Smith 1982) as a formalization of the neo-Darwinian concept of evolution via natural selection. It builds on the insight that many interactions between living beings can be considered to be games in the sense of game theory (GT) – every participant has something to win or to lose in the interaction, and the payoff of each participant can depend on the actions of all other participants. In the context of evolutionary biology, the payoff is an increase in fitness, where fitness is basically the expected number of offspring. According to the neo-Darwinian view on evolution, the units of natural selection are not primarily organisms but heritable traits of organisms. If the behavior of organisms, i.e., interactors, in a game-like situation is genetically determined, the strategies can be identified with gene configurations.

The evolutionary interpretation of GT is not confined to the biological context though. It is applicable to cultural evolution as well, were the transmission of strategies is achieved via imitation and learning rather than via DNA copying. Applied to language, EGT is thus a tool to model *conventionalization* formally, and this is of immediate relevance to the interface between pragmatics and grammar in the narrow sense.

3.2 Stability and dynamics

3.2.1 Evolutionary stability

Evolution is frequently conceptualized as a gradual progress towards more complexity and improved adaptation. Everybody has seen pictures displaying a linear ascent leading from algae over plants, fishes, dinosaurs, horses, and apes to Neanderthals and finally humans. Evolutionary biologists do not tire of pointing out that this picture is quite misleading. Darwinian evolution means a trajectory towards increased adaptation to the environment, proceeding in small steps. If a local maximum is reached, evolution is basically static. Change may occur if random variation (due to mutations) accumulate so that a population leaves its local optimum and ascends to another local optimum. Also, the fitness landscape itself may change as well – if the environment changes, the former optima may cease to be optimal. Most of the time biological evolution is macroscopically static though. Explaining stability is thus as important a goal for evolutionary theory as explaining change.

In the EGT setting, we are dealing with large populations of potential players. Each player is programmed for a certain strategy, and the members of the population play against each other very often under total random pairings. The payoffs of each encounter are accumulated as fitness, and the average number of offspring per individual is proportional to its accumulated fitness, while the birth rate and death rate are constant. Parents pass on their strategy to their offspring basically unchanged. Replication is to be thought of as asexual, i.e., each individual has exactly one parent. If a certain strategy yields on average a payoff that is higher than the population average, its replication rate will be higher than average and its proportion within the overall population increases, while strategies with a less-than-average expected payoff decrease in frequency. A strategy mix is stable under replication if the relative proportions of the different strategies within the population do not change under replication.

Occasionally replication is unfaithful though, and an offspring is programmed for a different strategy than its parent. If the mutant has a higher expected payoff (in games against members of the incumbent population) than the average of the incumbent population itself, the mutation will eventually spread and possibly drive the incumbent strategies to extinction. For this to happen, the initial number of mutants may be arbitrarily small.[8] Conversely, if the mutant does worse than the average incumbent, it will be wiped out and the incumbent strategy mix prevails.

A strategy mix is *evolutionarily stable* if it is resistant against the invasion of small proportions of mutant strategies. In other words, an evolutionarily stable strategy mix has an *invasion barrier*. If the number of mutant strategies is lower than this barrier, the incumbent strategy mix prevails, while invasions of higher numbers of mutants might still be successful.

In the metaphor used here, every player is programmed for a certain strategy, but a population can be mixed and comprise several strategies. Instead we may assume that all individuals are identically programmed, but this program is non-deterministic and plays different strategies according to some probability distribution (which corresponds to the relative frequencies of the pure strategies in the first conceptualization). Following the terminology from section 1, we call such non-deterministic strategies *mixed strategies*. For the purposes of the evolutionary dynamics of populations, the two models are equivalent. It is standard in EGT to talk of an *evolutionarily stable strategy*, where a strategy can be mixed, instead of an evolutionarily stable strategy mix. We will follow this terminology henceforth.

The notion of an evolutionarily stable strategy can be generalized to sets of strategies. A set of strategies A is stationary if a population where all individuals play a strategy from A will never leave A unless mutations occur. A set of strategies is evolutionarily stable if it is resistant against small amounts of non-A mutants. Especially interesting are minimal evolutionarily stable sets, i.e., evolutionarily stable sets which have no evolutionarily stable proper subsets. If the level of mutation is sufficiently small, each population will approach such a minimal evolutionarily stable set.

Maynard Smith (1982) gives a static characterization of evolutionarily stable strategies (ESS), abstracting away from the precise trajectories[9] of a population. It turns out that the notion of an ESS is strongly related to the rationalistic notions of a Nash equilibrium (NE) that was introduced earlier, and its stronger version of a strict Nash equilibrium (SNE). At the present point, we will focus on *symmetric* games where both players have the same strategies at their disposal, and we only consider profiles where both players play the same strategy. (The distinction between symmetric and asymmetric games will be discussed more thoroughly in the next subsection.)

With these adjustments, the definitions from section 1 can be rewritten as

- s is a **Nash Equilibrium** iff $u(s, s) \geq u(s, t)$ for all strategies t.
- s is a **Strict Nash Equilibrium** iff $u(s, s) > u(s, t)$ for all strategies t with $s \neq t$.

Are NEs always evolutionarily stable? Consider the well-known zero-sum game *Rock-Paper-Scissors* (RPS). The two players each have to choose between the three strategies R (rock), P (paper), and S (scissors). The rules are that R wins over S, S wins over P, and P wins over R. If both players play the same strategy, the result is a tie. A corresponding utility matrix would be as in Table 1.15. This game has exactly one NE. It is the mixed strategy $s*$ where one plays each pure strategy with a probability of $1/3$. If my opponent plays $s*$, my expected utility is 0, no matter what kind of strategy I play, because the probability of winning, losing, or a tie are equal. So every strategy is a best response to $s*$. On the other hand, if the probabilities of the strategies of my opponent are unequal, then my best response is always to play one of the pure strategies that win against the most probable of his actions. No strategy wins against itself; thus no other strategy can be a best response to itself. $s*$ is the unique NE .

Table 1.15: Utility matrix for Rock-Paper-Scissors

	R	P	S
R	0	-1	1
P	1	0	-1
S	-1	1	0

Is it evolutionarily stable? Suppose a population consists of equal parts of R, P, and S players, and they play against each other in random pairings. Then the players of each strategy have the same average utility, 0. If the number of offspring of each individual is positively correlated with its accumulated utility, there will be equally many individuals of each strategy in the next generation again, and the same in the second generation ad infinitum. $s*$ is a steady state. However, Maynard Smith's notion of evolutionary stability is stronger. An ESS should not only be stationary, but it should also be robust against mutations. Now suppose in a population as described above, some small proportion of the offspring of P-players are mutants and become S-players. Then the proportion of P-players in the next generation is slightly less than $\frac{1}{3}$, and the share of S-players exceeds $\frac{1}{3}$. So we have:

$$p(S) > p(R) > p(P)$$

This means that R-players will have an average utility that is slightly higher than 0 (because they win more against S and lose less against P). Likewise, S-players are at disadvantage because they win less than $\frac{1}{3}$ of the time (against P) but lose $\frac{1}{3}$ of the time (against R). So one generation later, the configuration is:

$$p(R) > p(P) > p(S)$$

By an analogous argument, the next generation will have the configuration:

$$p(P) > p(S) > p(R)$$

etc. After the mutation, the population has entered a circular trajectory, without ever approaching the stationary state $s*$ again without further mutations.

So not every NE is an ESS. The converse does hold though. Suppose a strategy s were not a NE . Then there would be a strategy t with $u(t, s) > u(s, s)$. This means that a t-mutant in a homogeneous s-population would achieve a higher average utility than the incumbents and thus spread. This may lead to the eventual extinction of s, a mixed equilibrium or a circular trajectory, but the pure s-population is never restored. Hence s is not an ESS. By contraposition we conclude that each ESS is a NE .

Can we identify ESSs with strict Nash equilibria (SNEs)? Not quite. Imagine a population of pigeons which come in two variants. A-pigeons have a perfect sense of orientation and can always find their way. B-pigeons have no sense of orientation at all. Suppose that pigeons always fly in pairs. There is no big disadvantage of being a B if your partner is of type A because he can lead the way. Likewise, it is of no disadvantage to have a B-partner if you are an A because you can lead the way yourself. (Let us assume for simplicity that leading the way has neither costs nor benefits.) However, a pair of B-individuals has a big disadvantage because it cannot find its way. Sometimes these pairs get lost and starve before they can reproduce. This corresponds to the utility matrix in Table 1.16 on the facing page. A is a NE, but not an SNE, because $u(B, A) = u(A, A)$. Now

Table 1.16: Utility matrix of the pigeon orientation game

	A	B
A	1	1
B	1	0

imagine that a homogeneous A-population is invaded by a small group of B-mutants. In a predominantly A-population, these invaders fare as well as the incumbents. However, there is a certain probability that a mutant goes on a journey with another B-mutant. Then both are in danger. Hence, sooner or later B-mutants will approach extinction because they cannot interact very well with their peers. More formally, suppose the proportions of A and B in the populations are $1 - \varepsilon$ and ε respectively. Then the average utility of A is 1, while the average utility of B is only $1 - \varepsilon$. Hence the A-subpopulation will grow faster than the B-subpopulation, and the share of B-individuals converges towards 0.

Another way to look at this scenario is this: B-invaders cannot spread in a homogeneous A-population, but A-invaders can successfully invade a B-population because $u(A, B) > u(B, B)$. Hence A is immune against B-mutants, even though A is only a non-strict Nash equilibrium.

If a strategy is immune against any kind of mutants in this sense, it is evolutionarily stable. The necessary and sufficient condition for evolutionary stability are (according to Maynard Smith 1982):

Definition 6 (Evolutionarily Stable Strategy) *s is an* Evolutionarily Stable Strategy *iff*

1 $u(s, s) \geq u(t, s)$ for all t, and

2 if $u(s, s) = u(t, s)$ for some $t \neq s$, then $u(s, t) > u(t, t)$.

The first clause requires an ESS to be a NE. The second clause says that if a t-mutation can survive in an s-population, s must be able to successfully invade any t-population for s to be evolutionarily stable.

From the definition it follows immediately that each SNE is an ESS. So we have the inclusion relation

Strict Nash Equilibria $\subset$ Evolutionarily Stable Strategies $\subset$ Nash Equilibria

Both inclusions are strict. The strategy A in the pigeon orientation game is evolutionarily stable without being a strict Nash equilibrium, and in Rock-Paper-Scissors, the mixed strategy to play each pure strategy with probability $\frac{1}{3}$ is a Nash equilibrium without being evolutionarily stable.

3.2.2 The replicator dynamics

The considerations that lead to the notion of an ESS are fairly general. They rest on three crucial assumptions:

1 Populations are (practically) infinite.

2 Each pair of individuals is equally likely to interact.

3 The expected number of offspring of an individual (i.e., its fitness in the Darwinian sense) is monotonically related to its average utility.

The assumption of infinity is crucial for two reasons. First, individuals usually do not interact with themselves under most interpretations of EGT . Thus, in a finite population, the probability to interact with a player using the same strategy as oneself would be less than the share of this strategy in the overall population. If the population is infinite, this discrepancy disappears. Second, in a finite population the average utility of players of a given strategy converges towards its expected value, but it need not be identical to it. This introduces a stochastic component. While this kind of stochastic EGT is a lively sub-branch of EGT (see below), the standard interpretation of EGT assumes deterministic evolution. In an infinite population, the average utility coincides with the expected utility .

As mentioned before, the evolutionary interpretation of GT interprets utilities as fitness. The notion of an ESS makes the weaker assumption that there is just a positive correlation between utility and fitness – a higher utility translates into more expected offspring, but this relation need not be linear. This is important for applications of EGT to cultural evolution, where replication proceeds via learning and imitation, and utilities correspond to social impact. There might be independent measures for utility that influence fitness without being identical to it.

Nevertheless it is often helpful to look at a particular population dynamics to sharpen one's intuition about the evolutionary behavior of a game. Also, in games such as Rock-Paper-Scissors, a lot of interesting things can be said about their evolution even though they have no stable states at all. Therefore we will discuss one particular evolutionary game dynamics in some detail.

The easiest way to relate utility and fitness in a monotonic way is of course just to identify them. So let us assume that the average utility of an individual equals its expected number of offspring. Let us say that there are n strategies $s_1, \ldots, s_n$. The amount of individuals playing strategy i is written as N_i. The relative frequency of strategy s_i, i.e., N_i/N, is written as x_i for short. (Note that x is a probability distribution, i.e. $\sum_j x_j = 1$.) We abbreviate the expected utility of strategy s_i, $\sum_{j=1}^{n} x_j u(i,j)$, as $\tilde{u}_i$, and the population average of the expected utility, $\sum_{i=1}^{n} x_i \tilde{u}_i$, as $\tilde{u}$.

If the population size N goes towards infinity, the development of the relative abundance of the different strategies within the population converges towards a deterministic dynamics that can be described by the following differential equation:

$$\frac{dx_i}{dt} \quad = \quad x_i(\tilde{u}_i - \tilde{u})$$

This equation is called the *replicator dynamics*. It was first introduced in Taylor and Jonker (1978). It is worth a closer examination. It says that the reproductive success of strategy s_i depends on two factors. First, there is the relative abundance of s_i itself, x_i. The more individuals in the current population are of type s_i, the more likely it is that there will be offspring of this type. The interesting part is the second factor, the *differential utility*. If $\tilde{u}_i = \tilde{u}$, this means that strategy s_i does exactly as well as the population average. In this case the two terms cancel each other out, and $\frac{dx_i}{dt} = 0$. This means that s_i's share of the total population remains constant. If, however, $\tilde{u}_i > \tilde{u}$, s_i does better than average, and it increases its share. Likewise, a strategy s_i with a less-than-average performance, i.e., $\tilde{u}_i < \tilde{u}$, loses ground.

Intuitively, evolutionary stability means a state is (a) stationary and (b) immune against the invasion of small numbers of mutations. This can directly be translated into dynamic notions. To require that a state is stationary amounts to saying that the relative frequencies of the different strategies within the population do not change over time. In other words, the vector x is stationary iff for all i:

$$\frac{dx_i}{dt} = 0$$

This is the case if either $x_i = 0$ or $\tilde{u}_i = \tilde{u}$ for all i.

Robustness against small amounts of mutation means that there is an environment of x such that all trajectories leading through this environment actually converge towards x. In the jargon of dynamic systems, x is then *asymptotically stable* or an *attractor*. It can be shown that a (possibly mixed) strategy is an ESS if and only if it is asymptotically stable under the replicator dynamics.

The replicator dynamics enables us to display the evolutionary behavior of a game graphically. This has a considerable heuristic value. There are basically two techniques for this. First, it is possible to depict *time series* in a Cartesian coordinate system. The time is mapped to the x-axis, while the y-axis corresponds to the relative frequency of some strategy. For some sample of initial conditions, the development of the relative frequencies over time is plotted as a function of the time variable. In a two-strategy game like the pigeon orientation scenario discussed above, this is sufficient to exhaustively display the dynamics of the system because the relative frequencies of the two strategies always sum up to 1. The left hand graphic in Figure 1.7 gives a few sample time series for the pigeon game. Here the y-axis corresponds to the relative frequency of the A-population. It is plainly obvious that the state where 100% of the population are of type A is in fact an attractor.

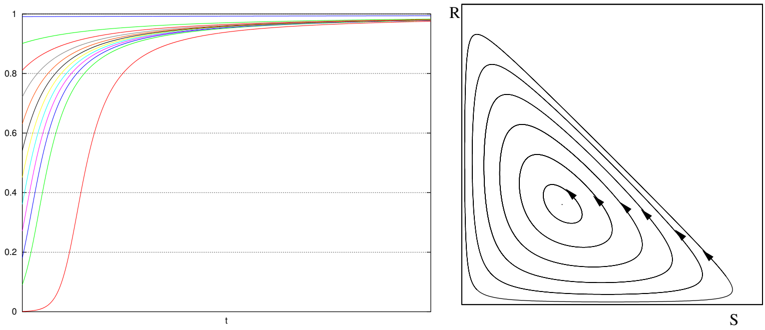


Figure 1.7: Replicator dynamics of the pigeon orientation game (left) and the rock-paper-scissor game (right)

Another option to graphically display the replicator dynamics of some game is to suppress the time dimension and instead plot possible *orbits* of the system. Here both axes correspond to relative frequencies of some strategies. So each state of the population corresponds to some point in the coordinate system. If there are at most two independent variables to consider – as in a symmetric three-strategy game like RPS – there is actually a 1-1 map between points and states. Under the replicator dynamics, populations evolve continuously. This corresponds to contiguous paths in the graph. The right hand graphic in Figure 1.7 shows some orbits of RPS. We plotted the frequencies of the "rock" strategy and the "scissors" strategy against the y-axis and the x-axis respectively. The sum of their frequencies never exceeds 1. This is why the whole action happens in the lower left corner of the square. The relative frequency of "paper" is uniquely determined by the two other strategies and is thus no independent variable.

The circular nature of this dynamics that we informally uncovered above is clearly discerned. One can also easily see "with the bare eye" that this game has no attractor, i.e., no ESS.

3.2.3 Asymmetric games

So far we considered *symmetric* games in this section. Formally, a game is symmetric iff the two players have the same set of strategies to choose from, and the utility does not depend on the position of the players. If u_1 is the utility matrix for the row player, and u_2 of the column player, then the game is symmetric iff both matrices are square (have the same number of rows and columns), and

$$u_1(i, j) = u_2(j, i)$$

There are various scenarios where these assumptions are inappropriate. In many types of interaction, the participants assume certain roles. In contests over a territory, it makes a difference who is the incumbent and who the intruder. In economic interaction, buyer and seller have different options at their disposal. Likewise in linguistic interaction you are the speaker or the hearer. The last example illustrates that it is possible for the same individual to assume either role on different occasions. If this is not possible, we are effectively dealing with two disjoint populations, like predators and prey or females and males in biology, haves and have-nots in economics, and adults and infants in language acquisition (in the latter case infants later become adults, but these stages can be considered different games).

The dynamic behavior of asymmetric games differs markedly from symmetric ones. The ultimate reason for this is that in a symmetric game, an individual can quasi play against itself (or against a clone of itself), while this is impossible in asymmetric games. The well-studied game "Hawks and Doves" may serve to illustrate this point. Imagine a population where the members have frequent disputes over some essential resource (food, territory, mates, whatever). There are two strategies to deal with a conflict. The aggressive type (the "hawks") will never give in. If two hawks come in conflict, they fight it out until one of them dies. The other one gets the resource. The doves, on the contrary, embark upon a lengthy ritualized dispute until one of them is tired of it and gives in. If a hawk and a dove meet, the dove gives in right away and the hawk gets the resource without any effort. There are no other strategies.

A possible utility matrix for this game is given in Table 1.17.

Table 1.17: Hawks and Doves

	H	D
H	1	7
D	2	3

Getting the disputed resource without effort has a survival value of 7.

Only a hawk meeting a dove is as lucky. Not getting the resource at all without a fight enables the loser to look out for a replacement. This is the fate of a dove meeting a hawk. Let's say this has a utility of 2. Dying in a fight over the resource leads to an expected number of 0 offspring, and a serious fight is also costly for the survivor. Let us say the average utility of a hawk meeting a hawk is 1. A dove meeting another dove will get the contested resource in one out of two occasions on average, but the lengthy ritualistic contest comes with a modest cost too, so the utility of a dove meeting a dove could be 3.

It is important to notice that the best response to a hawk is being a dove and vice versa. So neither of the two pure strategies is an NE. However, we also consider mixed strategies where either the population is mixed, or each individual plays either strategy with a certain probability. Under these circumstances, the game has an ESS. If the probability of behaving like a hawk is 80% and of being a dove 20%, both strategies achieve an expected utility of 2.2. As the reader may convince herself, this mixed strategy does in fact fulfill the conditions for an ESS. The replicator dynamics is given in Figure 1.8. Here the y-axis represents the proportion of hawks in a population. If the proportion of hawks exceeds the critical 80%, doves have an advantage and will spread, and vice versa. This changes dramatically if the same game is construed as an asymmetric game. Imagine the same situation as before, but now we are dealing with two closely related but different species. The two species are reproductively isolated, but they compete for the same ecological niche. Both species come in the hawkish and the dovish variant. Contests only take place between individuals from different species. Now suppose the first species, call it A, consists almost exclusively of the hawkish type. Under symmetric conditions, this would mean that the hawks mostly encounter another hawk, doves are better off on average, and

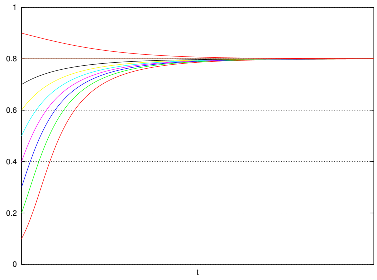


Figure 1.8: Symmetric Hawk-and-Dove game

therefore evolution works in favor of the doves. Things are different in the asymmetric situation. If A consists mainly of hawks, this supports the doves in the other species, B. So the proportion of doves in B will increase. This in turn reinforces the dominance of hawks in A. Likewise, a dominantly dovish A-population helps the hawks in B. The tendency always works in favor of a purely hawkish population in the one species and a purely dovish population in the other one.

Figure 1.9 graphically displays this situation. Here we use a third technique for visualizing a dynamics, a direction field. Each point in the plain corresponds to one state of the system. Here the x-coordinate gives the proportion of hawks in A, and the y-coordinate the proportion of hawks in B. Each arrow indicates in which direction the system is moving if it is in the state corresponding to the origin of the arrow. The length of the arrow indicates the velocity of the change. If you always follow the direction of the arrows, you get an orbit. Direction fields are especially useful to display systems with two independent variables, like the two-population game considered here. The system has two attractor states, the upper left and the lower right corner. They correspond to a purely hawkish population in one species and 100% doves in the other. If both populations have the critical 8:2 ratio of hawks:doves that was stable in the symmetric scenario, the system is also stationary. But this is not an attractor state because all points in the environment of this point are pulled away from it rather than being attracted to it.

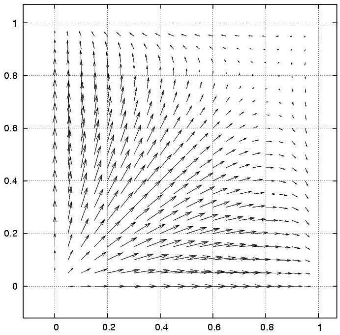


Figure 1.9: Asymmetric Hawk-and-Dove game

It is possible to capture the stability properties of asymmetric games in a way that is similar to the symmetric case. Actually the situation is even easier in the asymmetric case. Recall that the definition of a symmetric ESS was complicated by the consideration that mutants may encounter other mutants. In a two-population game, this is impossible. In a one-population role game, this might happen. However, minimal mutations only affect strategies in one of the two roles. If somebody minimally changes his grammatical preferences as a speaker, say, his interpretive preferences need not be affected by this.[10] So while a mutant might interact with its clone, it will never occur that a mutant strategy interacts with itself, because, by definition, the two strategy sets are distinct. So, the second clause of the definition of ESS doesn't matter.

To deal with asymmetric games, we have to use the more elaborate conceptual framework from section 1 again. Since the strategy sets of the two players (roles, populations) are distinct, the utility matrices are distinct as well. In a game between m and n strategies, the utility function of the first player is defined by an $m \times n$ matrix, call it u_A, and an $n \times m$ matrix u_B for the second player. An asymmetric Nash equilibrium is now a profile – a pair – of strategies, one for each population/role, such that each component is the best response to the other component. Likewise, a SNE is a pair of strategies where each one is the unique best response to the other.

Now if the second clause in the definition of a symmetric ESS plays no role here, does this mean that only the first clause matters? In other words, are all and only the NEs evolutionarily stable in the asymmetric case? Not quite. Suppose a situation as before, but now species A consists of three variants instead of two. The first two are both aggressive, and they both get the same, hawkish utility. Also, individuals from B get the same utility from interacting with either of the two types of hawks in A. The third A-strategy are still the doves. Now suppose that A consists exclusively of hawks of the first type, and B only of doves. Then the system is in a NE, since both hawk strategies are the best response to the doves in B, and for a B-player, being a dove is the best response to either hawk-strategy. If this A-population is invaded by a mutant of the second hawkish type, the mutants are exactly as fit as the incumbents. They will neither spread nor be extinguished. (Biologists call this phenomenon *drift* – change that has no impact for survival fitness and is driven by pure chance.) In this scenario, the system is in a (non-strict) NE, but it is not evolutionarily stable.

A *strict* NE is always evolutionarily stable though, and it can be shown (Selten 1980) that:

In asymmetric games, a configuration is an ESS iff it is a SNE .

It is a noteworthy fact about asymmetric games that ESSs are always pure in the sense that both populations play one particular strategy with 100% probability. This does not imply though that asymmetric games always settle in a pure state. Not every asymmetric game has an ESS. The asymmetric version of rock-paper-scissors, for instance, shows the same kind of cyclic dynamics as the symmetric variant.

As in the symmetric case, this characterization of evolutionary stability is completely general and holds for all utility monotonic dynamics. Again, the simplest instance of such a dynamic is the replicator dynamic. Here a state is characterized by two probability vectors, x and y. They represent the probabilities of the different strategies in the two populations or roles. The differential equation describing the replicator dynamics applies to multi-population games as well. The only difference is that the expected utility of a player from one population is calculated by averaging over the strategies for the other population.

3.3 EGT and language

Language is first and foremost a means for communication. As a side effect of communication, linguistic knowledge is transmitted between the communicators. This is most obvious in language acquisition, but learning never stops, and adult speakers of the same language exert a certain influence on each other's linguistic habits as well. This makes natural language an interactive and self-replicative system. Hence EGT is a promising analytical tool for the study of linguistic phenomena. Let us start this subsection with a few general remarks.

To give an EGT formalization – or an evolutionary conceptualization in general – of a particular empirical phenomenon, various issues have to be addressed in advance. What is replication in the domain in question? What are the units of replication? Is replication faithful, and if so, which features are constant under replication? What factors influence reproductive success (= fitness)? What kind of variation exists, and how does it interact with replication?

There are various aspects of natural language that are subject to replication, variation and selection, on various timescales that range from minutes (single discourse) to millennia (language related aspects of biological evolution). We will focus on cultural (as opposed to biological) evolution on short time scales, but we will briefly discuss the more general picture.

The most obvious mode of linguistic self-replication is first language acquisition. Before this can take effect, the biological preconditions for language acquisition and use have to be given, ranging from the physiology of the ear and the vocal tract to the necessary cognitive abilities. The biological language faculty is replicated in biological reproduction. It seems obvious that the ability to communicate does increase survival chances and social standing and thus promotes biological fitness, but only at a first glance.

Sharing information usually benefits the receiver more than the sender because information arguably increases fitness. Sharing information with others increases the fitness of the others and thus reduces the own *differential* fitness. Standard EGT predicts this kind of altruistic behavior to be evolutionarily unstable. Here is a crude formalization in terms of an asymmetric game between sender and receiver. The sender has a choice between sharing information ("T" for "talkative") or keeping information for himself ("S" for "silent"). The (potential) receiver has the options of paying attention and trying to decode the messages of the sender ("A" for "attention") or to ignore ("I") the sender. Let us say that sharing information does have a certain benefit for the sender because it may serve to manipulate the receiver. On the other hand, sending a signal comes with an effort and may draw the attention of predators. For the sake of the argument, we assume that the costs and benefits are roughly equally distributed *given that the receiver pays attention*. If the receiver ignores the message, it is disadvantageous for the sender to be talkative. For the receiver, it pays to pay attention if the sender actually sends. Then the listener benefits most. If the sender is silent, it is of disadvantage for the listener to pay attention because attention is a precious resource that could have been spent in a more useful way otherwise. Sample utilities that mirror these assumed preferences are given in Table 1.18.

Table 1.18: The utility of communication

	A	I
T	(1 ; 2)	(0 ; 1)
S	(1 ; 0)	(1 ; 1)

The game has exactly one ESS, namely the combination of "S" and "I". (As the careful reader probably already figured out for herself, a cell is an ESS, i.e., a strict Nash equilibrium, if its first number is the unique maximum in its column and the second one the unique maximum in its row.) This result might seem surprising. The receiver would actually be better off if the two parties would settle at (T,A). This would be of no disadvantage for the sender. Since the sender does not compete with the receiver for resources (we are talking about an asymmetric game), he could actually afford to be generous and grant the receiver the possible gain. Here the predictions of standard EGT seem to be at odds with the empirical observations.[11] The evolutionary basis for communication, and for cooperation in general, is an active area of research in EGT , and there are various possible routes that have been proposed.

First, the formalization that we gave here may be just wrong, and communication is in fact beneficial for both parties. While this is certainly true for humans living in human societies, this still raises the questions how these societies could have evolved in the first place.

A more interesting approach goes under the name of the *handicap principle*. The name was coined by Zahavi (1975) to describe certain patterns of seemingly self-destructive communication in the animal kingdom. A good example is what he calls the "stotting" behavior of gazelles:

> We start with a scene of a gazelle resting or grazing in the desert. It is nearly invisible; the color of its coat bends well with the desert landscape. One would expect the gazelle to freeze or crouch and do its utmost to avoid being seen. But no: it rises, barks, and thumps the ground with its forefeet, all the while watching the wolf. [...] Why does the gazelle reveal itself to a predator that might not otherwise spot it? Why does it waste time and energy jumping up and down (*stotting*) instead of running away as fast as it can? The gazelle is signaling to the predator that it has seen it; by "wasting" time and jumping high in the air rather than bounding away, it demonstrates in a reliable way that it is able to outrun the wolf. The wolf, upon learning that it has lost its chance to surprise its prey, and that this gazelle is in top-top physical shape, may decide to move on to another area; or it may decide to look for more promising prey. (from Zahavi and Zahavi 1997, xiii-xiv)

Actually, the best response of the predator is to call the bluff occasionally, often enough to deter cheaters, but not too often. Under these conditions, the self-inflicted handicap of the (fast) gazelle is in fact evolutionarily stable.

The crucial insight here is that truthful communication can be evolutionarily stable if lying is more costly than communicating the truth. A slow gazelle could try to use stotting as well to discourage a lion from hunting it, but this would be risky if the lion occasionally calls the bluff. The expected costs of such a strategy are thus higher than the costs of running away immediately. In communication among humans, there are various ways in which lying might be more costly than telling (or communicating) the truth. To take an example from economics, driving a Rolls Royce communicates "I am rich" because for a poor man, the costs of buying and maintaining such an expensive car outweigh its benefits while a rich man can afford them. Here producing the signal as such is costly. In linguistic communication, lying comes with the social risk of being found out, so in many cases telling the truth is more beneficial than lying.

The idea of the handicap principle as an evolutionary basis for communication has inspired a plethora of research in biology and economics. Van

Rooij (2003) uses it to give a game theoretic explanation of politeness as a pragmatic phenomenon.

A third hypothesis rejects the assumption of standard EGT that all individuals interact with equal probability. When I increase the fitness of my kin, I thereby increase the chances for replication of my own gene pool, even if it should be to my own disadvantage. Recall that utility in EGT does not mean the reproductive success of an individual but of a *strategy*, and strategies correspond to heritable traits in biology. A heritable trait for altruism might thus have a high expected utility provided its carrier preferably interacts with other carriers of this trait. Biologists call this model *kin selection*. There are various modifications of EGT that give up the assumption of random pairing. Space does not permit us to go into any detail here. However, refinements of EGT where a player is more likely to interact with other individuals of its own type often predict cooperative or even altruistic behavior to be evolutionarily stable even if it not an ESS according to Maynard Smith's criteria.

Natural languages are not passed on via biological but via cultural transmission. First language acquisition is thus a qualitatively different mode of replication. Most applications of evolutionary thinking in linguistics focus on the ensuing acquisition driven dynamics. It is an important aspect in understanding language change on a historical timescale of decades and centuries.

It is important to notice that there is a qualitative difference between Darwinian evolution and the dynamics that results from iterated learning (in the sense of iterated first language acquisition). In Darwinian evolution, replication is almost always faithful. Variation is the result of occasional unfaithful replication, a rare and essentially random event. Theories that attempt to understand language change via iterated language acquisition stress the fact though that here, replication can be unfaithful in a systematic way. The work of Martin Nowak and his co-workers (see for instance Nowak et al. 2002) is a good representative of this approach. They assume that an infant that grows up in a community of speakers of some language L_1 might acquire another language L_2 with a certain probability. This means that those languages will spread in a population that (a) are likely targets of acquisition for children that are exposed to other languages, and (b) are likely to be acquired faithfully themselves. This approach thus conceptualizes language change as a Markov process[12] rather than evolution through natural selection. Markov processes and natural selection of course do not exclude each other. Nowak's differential equation describing the language acquisition dynamics actually consists of a basically game theoretical natural selection component (pertaining to the functionality of language) and a (learning oriented) Markov component.

Language is also replicated on a much shorter time scale, just via being used. The difference between acquisition based and usage based replication can be illustrated by looking at the development of the vocabulary of some language. There are various ways how a new word can enter a language – morphological compounding, borrowing from other languages, lexicalization of names, coinage of acronyms, and what have you. Once a word is part of a language, it is gradually adapted to this language, i.e., it acquires a regular morphological paradigm, its pronunciation is nativized etc. The process of establishing a new word is predominantly driven by mature (adult or adolescent) language users, not by infants. Somebody introduces the new word, and people start imitating it. Whether the new coinage catches on depends on whether there is a need for this word, whether it fills a social function (like distinguishing the own social group from other groups), whether the persons who already use have a high social prestige etc.

Since the work of Labov (see for instance Labov 1972) functionally oriented linguists have repeatedly pointed out that grammatical change actually follows a similar pattern. The main agents of language change, they argue, are mature language users rather than children. Not just the vocabulary is plastic and changes via language use but all kinds of linguistic variables like syntactic constructions, phones, morphological devices, interpretational preferences etc. Imitation plays a crucial part here, and imitation is of course a kind of replication. Unlike in biological replication, the usage of a certain word or construction can usually not be traced back to a unique model or pair of models that spawn the token in question. Rather, every previous usage of this linguistic item that the user took notice of shares a certain fraction of "parenthood". Recall though that the basic units of evolution in EGT are not individuals but strategies, and evolution is about the relative frequency of strategies. If there is a causal relation between the abundance of a certain linguistic variant at a given point in time and its abundance at a later point, we can consider this a kind of faithful replication. Also, replication is almost but not absolutely faithful. This leads to a certain degree of variation. Competing variants of a linguistic item differ in their likelihood to be imitated – this corresponds to fitness and thus to natural selection. The usage based dynamics of language use has all aspects that are required for a modeling in terms of EGT .

In the linguistic examples that we will discuss further on, we will assume the latter notion of linguistic evolution.

3.4 Pragmatics and EGT

In this subsection we will go through a couple of examples that demonstrate how EGT can be used to explain high level linguistic notions like pragmatic preferences or functional pressure. For more detailed accounts of linguistic phenomena using EGT, the reader is referred to Jäger (2004) and van Rooij (2004).

3.4.1 Partial blocking

If there are two comparable expressions in a language such that the first is strictly more specific than the second, there is a tendency to reserve the more general expression for situations where the more specific one is not applicable. A standard example is the opposition between "many" and "all". If I say that many students came to the guest lecture, it is usually understood that not all students came. There is a straightforward rationalistic explanation for this in terms of conversational maxims: the speaker should be as specific as possible. If the speaker uses "many", the hearer can conclude that the usage of "all" would have been inappropriate. This conclusion is strictly speaking not valid though – it is also possible that the speaker just does not know whether all students came or whether a few were missing.

A similar pattern can be found in conventionalized form in the organization of the lexicon. If a regular morphological derivation and a simplex word compete, the complex word is usually reserved for cases where the simplex is not applicable. For instance, the compositional meaning of the English noun "cutter" is just *someone or something that cuts*. A knife is an instrument for cutting, but still you cannot call a knife a "cutter". The latter word is reserved for non-prototypical cutting instruments.

Let us consider the latter example more closely. We assume that the literal meaning of "cutter" is a concept CUTTER′ and the literal meaning of "knife" a concept KNIFE′ such that every knife is a cutter but not vice versa, i.e.,

$$\text{KNIFE}' \subset \text{CUTTER}'$$

There are two basic strategies to use these two words, the *semantic* (S) and the *pragmatic* (P) strategy. Both come in two versions, a hearer strategy and a speaker strategy. A speaker using S will use "cutter" to refer to unspecified cutting instruments, and "knife" to refer to knives. To refer to a cutting instrument that is not a knife, this strategy either uses the explicit "cutter but not a knife", or, short but imprecise, also "cutter". A hearer using S will interpret every expression literally, i.e., "knife" means KNIFE′, "cutter" means CUTTER′, and "cutter but not a knife" means CUTTER′ − KNIFE′.

A speaker using P will reserve the word "cutter" for the concept CUTTER′ − KNIFE′. To express the general concept CUTTER′, this strategy has to resort to a more complex expression like "cutter or knife". Conversely, a hearer using P will interpret "cutter" as CUTTER′ − KNIFE′, "knife" as KNIFE′, and "cutter or knife" as CUTTER′.

So we are dealing with an asymmetric 2 × 2 game. What is the utility function? In EGT, utilities are interpreted as the expected number of offspring. In our linguistic interpretation this means that utilities express the likelihood of a strategy to be imitated. It is a difficult question to tease apart the factors that determine the utility of a linguistic item in this sense, and ultimately it has to be answered by psycholinguistic and sociolinguistic research. Since we have not undertaken this research so far, we will make up a utility function, using plausibility arguments.

We start with the hearer perspective. The main objective of the hearer in communication, let us assume, is to gain as much truthful information as possible. The utility of a proposition for the hearer is thus inversely proportional to its probability, *provided the proposition is true*. For the sake of simplicity, we only consider contexts where the nouns in question occur in upward entailing context. Therefore CUTTER′ has a lower information value than KNIFE′ or CUTTER′−KNIFE′. It seems also fair to assume that non-prototypical cutters are more rarely talked about than knives; thus the information value of KNIFE′ is lower than the one of CUTTER′−KNIFE′.

For concreteness, we make up some numbers. If i is the function that assigns a concept its information value, let us say that

$$\begin{aligned} i(\text{KNIFE}') &= 30 \\ i(\text{CUTTER}' - \text{KNIFE}') &= 40 \\ i(\text{CUTTER}') &= 20 \end{aligned}$$

The speaker wants to communicate information. Assuming only honest intentions, the information value that the hearer gains should also be part of the speaker's utility function. Furthermore, the speaker wants to minimize his effort. So as a second component of the speaker's utility function, we assume some complexity measure over expressions. A morphologically complex word like "cutter" is arguably more complex than a simple one like "knife", and syntactically complex phrases like "cutter or knife" or "cutter but not knife" are even more complex. The following stipulated values for the cost function take these considerations into account:

$$\begin{aligned} \mathit{cost}(\text{“knife”}) &= 1 \\ \mathit{cost}(\text{“cutter”}) &= 2 \\ \mathit{cost}(\text{“cutter or knife”}) &= 40 \\ \mathit{cost}(\text{“cutter but not knife”}) &= 45 \end{aligned}$$

These costs and benefits are to be weighted – everything depends on how often each of the candidate concepts is actually used. The most prototypical concept of the three is certainly KNIFE′, while the unspecific CUTTER′ is arguably rare. Let us say that, conditioned to all utterance situations in question, the probabilities that a speaker tries to communicate the respective concepts are

$$\begin{aligned} p(\text{KNIFE}') &= .7 \\ p(\text{CUTTER}' - \text{KNIFE}') &= .2 \\ p(\text{CUTTER}') &= .1 \end{aligned}$$

The utility of the speaker is then the difference between the average information value that he manages to communicate and the average costs that he has to afford. The utility of the hearer is just the average value of the correct information that is received. The precise values of these utilities finally depend on how often a speaker of the S-strategy actually uses the complex "cutter but not knife", and how often he uses the shorter "cutter". Let us assume for the sake of concreteness that he uses the short form in 60% of all times.

After some elementary calculations, this leads us to the following utility matrix. The speaker is assumed to be the row player and the hearer the column player. Both players receive the absolutely highest utility if both play P. This means perfect communication with minimal effort. All other combinations involve some kind of communication failure because the hearer occasionally interprets the speaker's use of "cutter" either too strongly or too weakly.

Table 1.19: Knife vs. cutter

	S	P
S	(23.86 ; 28.60)	(24.26 ; 29.00)
P	(23.40 ; 29.00)	(25.40 ; 31.00)

If both players start out with the semantic strategy, mutant hearers that use the pragmatic strategy will spread because they get the more specific interpretation CUTTER′−KNIFE′ right in all cases where the speaker prefers minimizing effort over being explicit. The mutants will get all cases wrong where the speaker meant CUTTER′ by using "cutter", but the advantage is greater. If the hearers employ the pragmatic strategy, speakers using their pragmatic strategy will start to spread now because they will have a higher chance to get their message across. The combination P/P is the only strict Nash equilibrium in the game and thus the only ESS.

Figure 1.10 gives the direction field of the corresponding replicator dynamics. The x-axis gives the proportion of the hearers that are P-players, and the y-axis corresponds to the speaker dimension. The structural properties of this game are very sensitive to the particular parameter values. For instance, if the informational value of the concept CUTTER′ were 25 instead of 20, the resulting utility matrix would come out as in Table 1.20. Here both S/S and P/P come out as evolutionarily stable. This result is not entirely unwelcome – there are plenty of examples where a specific term does not block a general term. If I refer to a certain dog as "this dog", I do not implicate that it is of no discernible breed like "German shepherd" or "Airedale terrier". The more general concept of a dog is useful enough to prevent blocking by more specific terms.

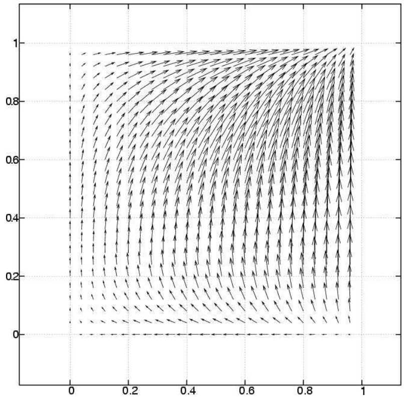


Figure 1.10: Partial blocking: replicator dynamics

Table 1.20: Knife vs. cutter, different parameter values

	S	P
S	$(24.96\ ;\ 29.70)$	$(24.26\ ;\ 29.00)$
P	$(23.40\ ;\ 30.00)$	$(25.90\ ;\ 31.50)$

3.4.2 Horn strategies

Real synonymy is rare in natural language – some people even doubt that it exists. Even if two expressions should have identical meanings according to the rules of compositional meaning interpretation, their actual interpretation is usually subtly differentiated. Larry Horn (see for instance Horn 1993) calls this phenomenon the *division of pragmatic labor*. This differentiation is not just random. Rather, the tendency is that the simpler of the two competing expressions is assigned to the prototypical instances of the common meaning, while the more complex expression is reserved for less prototypical situations. The following examples (taken from op. cit.) serve to illustrate this.

(8) a. John went to church/jail. (prototypical interpretation)

b. John went to the church/jail. (literal interpretation)

(9) a. I need a new driller/cooker.

b. I need a new drill/cook.

The example (8a) only has the non-literal meaning where John attended a religious service or was convicted and send to a prison respectively. The more complex (b) sentence literally means that he approaches the church (jail) as a pedestrian.

De-verbal nouns formed by the suffix *-er* can either be agentive or refer to instruments. So compositionally, a *driller* could be a person who drills or an instrument for drilling, and likewise for *cooker*. However, *drill* is lexicalized as a drilling instrument, and thus *driller* can only have the agentive meaning. For *cooker* it is the other way round: a *cook* is a person who cooks, and thus a *cooker* can only be an instrument. Arguably the concept of a person who cooks is a more natural concept than an instrument for cooking in our culture, and for drills and drillers it is the other way round. So in either case, the simpler form is restricted to the more prototypical meaning.

One might ask what "prototypical" exactly means here. The meaning of "going to church" for instance is actually more complex than the meaning of "going to the church" because the former invokes a lot of cultural background knowledge. It seems to make sense to us though to simply identify prototypicality with frequency. Those meanings that are most often communicated in ordinary conversations are most prototypical. We are not aware whether anybody carried out any quantitative studies on this subject, but simple Google searches show that for the mentioned examples, this seems to be a good hypothesis. The phrase "went to church" got 88,000 hits, against

13,500 for "went to the church". "I will marry you" occurs 5,980 times; "I am going to marry you" only 442 times. "A cook" has about 712,000 occurrences while "a cooker" has only about 25,000. (This crude method is not applicable to "drill" vs. "driller" because the former also has an additional meaning as in "military drill" which pops up very often.)

While queries at a search engine do not replace serious quantitative investigations, we take it to be a promising hypothesis that in case of a pragmatic competition, the less complex form tends to be restricted to the more frequent meaning and the more complex one to the less frequent interpretation. It is straightforward to formalize this setting in a game. The players are speaker and hearer. There are two meanings that can be communicated, m_0 and m_1, and they have two forms at their disposal, f_0 and f_1.

Each total function from meanings to forms is a speaker strategy, while hearer strategies are mappings from forms to meanings. There are four strategies for each player, as shown in Table 1.21 on the following page.

It is decided by nature which meaning the speaker has to communicate. The probability that nature chooses m_0 is higher than the probability of m_1. Furthermore, form f_0 is less complex than form f_1.

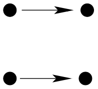
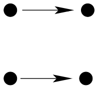
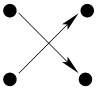
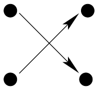
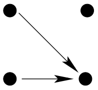
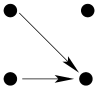
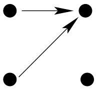
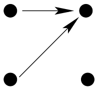
So far this is not different from the signaling games from section 2. However, we assume here that talk is not cheap. (For simplicity's sake, we identify both types and actions with meanings here.) The speaker has an interest in minimizing the complexity of the expression involved. One might argue that the hearer also has an interest in minimizing complexity. However, the hearer is confronted with a given form and has to make sense of it. He or she has no way to influence the complexity of that form or the associated meaning. Therefore there is no real point in making complexity part of the hearer's utility function.

To keep things simple, let us make up some concrete numbers. Let us say that the probability of m_1 is 75% and the probability of m_2 25%. The costs of f_1 and f_2 are 0.1 and 0.2 respectively. The unit is the reward for successful communication – so we assume that it is 10 times as important for the speaker to get the message across than to avoid the difference in costs between f_2 and f_1. We exclude strategies where the speaker does not say anything at all, so the minimum cost of 0.1 unit is unavoidable.

The utility of the hearer for a given pair of a hearer strategy and a speaker strategy is the average number of times that the meaning comes across correctly given the strategies and nature's probability distribution. Formally this means that

$$u_h(H, S) \quad = \quad \sum_m p_m \times \delta_m(S, H)$$

Table 1.21: Strategies in the Horn game

	Speaker			Hearer	
$S1$:	$m_0 \mapsto f_0$ $m_1 \mapsto f_1$		$H1$:	$f_0 \mapsto m_0$ $f_1 \mapsto m_1$	
$S2$:	$m_0 \mapsto f_1$ $m_1 \mapsto f_0$		$H2$:	$f_0 \mapsto m_1$ $f_1 \mapsto m_0$	
$S3$:	$m_0 \mapsto f_0$ $m_1 \mapsto f_0$		$H3$:	$f_0 \mapsto m_0$ $f_1 \mapsto m_0$	
$S4$:	$m_0 \mapsto f_1$ $m_1 \mapsto f_1$		$H4$:	$f_0 \mapsto m_1$ $f_1 \mapsto m_1$	

where the δ-function is defined as

$$\delta_m(S,H) = \begin{cases} 1 & \text{if} \quad H(S(m)) = m \\ 0 & \text{else} \end{cases}$$

The speaker shares the interest in communicating successfully, but he also wants to avoid costs. So his utility function comes out as

$$u_s(S,H) \quad = \quad \sum_m p_m \times (\delta_m(S,H) - \mathit{cost}(S(m)))$$

With the chosen numbers, this gives us the utility matrix in Table 1.22 on the next page. The first question that might come to mind is what negative utilities are supposed to mean in EGT. Utilities are the expected number of offspring – what is negative offspring? Recall though that if applied to cultural language evolution, the replicating individuals are utterances, and the

Table 1.22: Utility matrix of the Horn game

	H_1	H_2	H_3	H_4
S_1	$(.875\ ;\ 1.0)$	$(-.125\ ;\ 0.0)$	$(.625\ ;\ .75)$	$(.125\ ;\ .25)$
S_2	$(-.175\ ;\ 0.0)$	$(.825\ ;\ 1.0)$	$(.575\ ;\ .75)$	$(.25\ ;\ .075)$
S_3	$(.65\ ;\ .75)$	$(.15\ ;\ .25)$	$(.65\ ;\ .75)$	$(15\ ;\ .25)$
S_4	$(.05\ ;\ .25)$	$(.55\ ;\ .75)$	$(.55\ ;\ .75)$	$(.05\ ;\ .25)$

mode of replication is imitation. Here the utilities represent the difference in the absolute abundance of a certain strategy at a given point in time and at a later point. A negative utility thus simply means that the number of utterances generated by a certain strategy is absolutely declining.

Also, neither the replicator dynamics nor the locations of ESSs or Nash equilibria change if a constant amount is added to all utilities within a matrix. It is thus always possible to transform any given matrix into an equivalent one with only non-negative entries.

We are dealing with an asymmetric game. Here all and only the strict Nash equilibria are evolutionarily stable. There are two such stable states in the game at hand: (S_1, H_1) and (S_2, H_2). As the reader may verify, these are the two strategy configurations where both players use a 1-1 function, the hearer function is the inverse of the speaker function, and where thus communication always succeeds. EGT thus predicts the emergence of signaling conventions in the Lewisian sense.

It does not predict though that the "Horn strategy" (S_1, H_1) is in any way superior to the "anti-Horn strategy" (S_2, H_2) where the complex form is used for the frequent meaning. There are various reasons why the former strategy is somehow "dominant". First, it is *Pareto optimal* (recall the discussion of Pareto optimality on page 23). This means that for both players, the utility that they get if both play Horn is at least as high as in the other ESS where they both play anti-Horn. For the speaker Horn is absolutely preferable. Horn also *risk-dominates* anti-Horn. This means that if both players play Horn, either one would have to lose a lot by deviating unilaterally to anti-Horn, and this "risk" is at least as high as the inverse risk, i.e., the loss in utility from unilaterally deviating from the anti-Horn equilibrium. For the speaker, this domination is strict.

However, these considerations are based on a rationalistic conception of GT , and they are not directly applicable to EGT . There are two arguments for the domination of the Horn strategy that follow directly from the replicator dynamics.

- A population where all eight strategies are equally likely will converge towards a Horn strategy. Figure 1.11 gives the time series for all eight strategies if they all start at 25% probability. Note that the hearers first pass a stage where strategy H_3 is dominant. This is the strategy where the hearer always "guesses" the more frequent meaning – a good strategy as long as the speaker is unpredictable. Only after the speaker starts to clearly differentiate between the two meanings does H_1 begin to flourish.

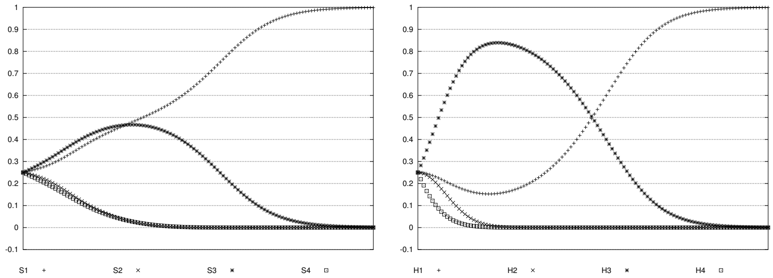


Figure 1.11: Time series of the Horn game

- While both Horn and anti-Horn are attractors under the replicator dynamics, the former has a much larger basin of attraction than the latter. We are not aware of a simple way of analytically calculating the ratio of the sizes of the two basins, but a numerical approximation revealed that the basin of attraction of the Horn strategy is about 20 times as large as the basin of attraction of the anti-Horn strategy.

The asymmetry between the two ESSs becomes even more apparent when the idealization of the population being infinite population is lifted. In the next section we will briefly explore the consequences of this.

3.5 All equilibria are stable, but some equilibria are more stable than others: Stochastic EGT

Let us now have a closer look at the modeling of mutations in EGT. Evolutionary stability means that a state is stationary and resistant against small amounts of mutations. This means that the replicator dynamics is tacitly assumed to be combined with a small stream of mutation from each strategy to each other strategy. The level of mutation is assumed to be constant. An evolutionarily stable state is a state that is an attractor in the combined dynamics and remains an attractor as the level of mutation converges towards zero.

The assumption that the level of mutation is constant and deterministic, though, is actually an artifact of the assumption that populations are infinite and time is continuous in standard EGT. Real populations are finite, and both games and mutations are discrete events in time. So a more fine-grained modeling should assume finite populations and discrete time. Now suppose that for each individual in a population, the probability to mutate towards the strategy s within one time unit is p, where p may be very small but still positive. If the population consists of n individuals, the chance that all individuals end up playing s at a given point in time is at least p^n, which may be extremely small but is still positive. By the same kind of reasoning, it follows that there is a positive probability for a finite population to jump from each state to each other state due to mutation (provided each strategy can be the target of mutation of each other strategy). More generally, in a finite population the stream of mutations is not constant but noisy and non-deterministic. Hence there are strictly speaking no evolutionarily stable strategies because every invasion barrier will eventually be overcome, no matter how low the average mutation probability or how high the barrier.[13]

If an asymmetric game has exactly two SNEs, A and B, in a finite population with mutations there is a positive probability p_{AB} that the system moves from A to B due to noisy mutation, and a probability p_{BA} for the reverse direction. If $p_{AB} > p_{BA}$, the former change will on average occur more often than the latter, and in the long run the population will spend more time in state B than in state A. Put differently, if such a system is observed at some arbitrary time, the probability that it is in state B is higher than that it is in A. The exact value of this probability converges towards $\frac{p_{AB}}{p_{AB}+p_{BA}}$ as time grows to infinity.

If the level of mutation gets smaller, both p_{AB} and p_{BA} get smaller, but at a different pace. p_{BA} approaches 0 much faster than p_{AB}, and thus $\frac{p_{AB}}{p_{AB}+p_{BA}}$ (and thus the probability of the system being in state B) converges to 1 as the mutation rate converges to 0. So while there is always a positive probability that the system is in state A, this probability can become arbitrarily small. A state is called *stochastically stable* if its probability converges to a value > 0 as the mutation rate approaches 0. In the described scenario, B would be the only stochastically stable state, while both A and B are evolutionarily stable. The notion of stochastic stability is a strengthening of the concept of evolutionary stability; every stochastically stable state is also evolutionarily stable,[14] but not the other way round.

We can apply these considerations to the equilibrium selection problem in the Horn game from the last subsection. Figure 1.12 shows the results of a simulation, using a stochastic dynamics in the described way.[15] The left hand figure shows the proportion of the Horn strategy S_1 and the figure on the right the anti-Horn strategy S_2. The other two speaker strategies remain close to zero. The development for the hearer strategies is pretty much synchronized. During the simulation, the system spent 67% of the time in a state with a predominant Horn strategy and only 26% with predominant anti-Horn (the remaining time are the transitions). This seems to indicate strongly that the Horn strategy is in fact the more probable one, which in turn indicates that it is the only stochastically stable state.

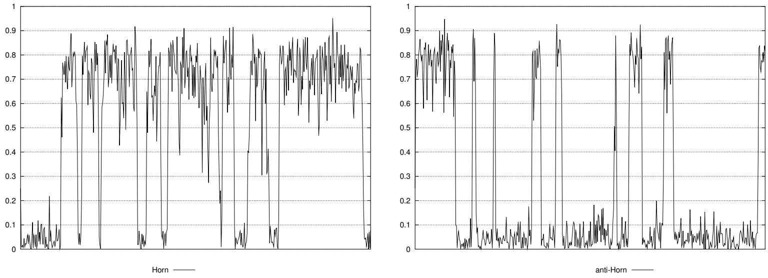


Figure 1.12: Simulation of the stochastic dynamics of the Horn game

The literature contains some general results about how to find the stochastically stable states of a system analytically, but they are all confined to 2×2 games. This renders them practically useless for linguistic applications because here, even in very abstract models like the Horn game, we deal with more than two strategies per player. For larger games, analytical solutions can only be found by studying the properties in question on a case by case basis. It would take us too far to discuss possible solution concepts here in detail (see for instance Young 1998 or Ellison 2000). We will just sketch such an analytical approach for the Horn game, which turns out to be comparatively well-behaved.

To check which of the two ESSs of the Horn game are stochastically stable, we have to compare the height of their invasion barriers. How many speakers must deviate from the Horn strategy such that even the smallest hearer mutation causes the system to leave the basin of attraction of this strategy and to move towards the anti-Horn strategy? And how many hearer-mutations would have this effect? The same questions have to be answered for the anti-Horn strategy, and the results to be compared.

Consider speaker deviations from the Horn strategy. It will only lead to an incentive for the hearer to deviate as well if H_1 is not the optimal response to the speaker strategy anymore. This will happen if at least 50% of all speakers deviate toward S_2, 66.7% deviate towards S_4, or some combination of such deviations. It is easy to see that the minimal amount of deviation having the effect in question is 50% deviating towards S_2.[16]

As for hearer deviation, it would take more than 52.5% mutants towards H_2 to create an incentive for the speaker to deviate towards S_2, and even about 54% of deviation towards H_4 to have the same effect. So the invasion barrier along the hearer dimension is 52.5%.

Now suppose the system is in the anti-Horn equilibrium. As far as hearer utilities are concerned, Horn and anti-Horn are completely symmetrical, and thus the invasion barrier for speaker mutants is again 50%. However, if more than 47.5% of all hearers deviate towards H_1, the speaker has an incentive to deviate towards S_1.

In sum, the invasion barriers of the Horn and of the anti-Horn strategy are 50% and 47.5% respectively. Therefore a "catastrophic" mutation from the latter to the former, though unlikely, is more likely than the reverse transition. This makes the Horn strategy the only stochastically stable state.

In this particular example, only two strategies for each player played a role in determining the stochastically stable state. The Horn game thus behaves effectively as a 2×2 game. In such games stochastic stability actually coincides with the rationalistic notion of "risk dominance" that was briefly discussed above. In the general case, it is possible though that a larger game has two ESSs, but there is a possible mutation from one equilibrium towards a third state (for instance a non-strict Nash equilibrium) that lies within the basin of attraction of the other ESS. The stochastic analysis of larger games has to be done on a case-by-case basis to take such complex structures into account.

In standard EGT, as well as in the version of stochastic EGT discussed here, the utility of an individual at each point in time is assumed to be exactly the average utility this individual would get if it played against a perfectly representative sample of the population. Vega-Redondo (1996) discusses another variant of stochastic EGT where this idealization is also given up. In this model, each individual plays a finite number of tournaments in each time step, and the gained utility – and thus the abundance of offspring – becomes a stochastic notion as well. He shows that this model sometimes leads to a different notion of stochastic stability than the one discussed here. A detailed discussion of this model would lead beyond the scope of this introduction though.

4 Overview

With this book we hope to attract the attention of researchers and students of linguistics and of the philosophy of language that are interested in pragmatics. We hope to convince those readers of the potential and the relevance of game theory for linguistic pragmatics, and for the understanding of language in general. Likewise, we hope to convince working game theorists from other fields that natural language is an exciting area of application of their theory.

Even though the roots of game theoretic pragmatics go back to the late sixties, it is still an emerging discipline. This makes the field diverse, and at the same time exciting and innovative. There is no agreement yet on a set of established ideas, concepts, and research questions, and in a sense, this is what makes the field so attractive for researchers from different backgrounds. In this volume, we hope to give a snapshot of the current state of this budding discipline.

Lewis (1969) introduced signaling games for the study of linguistic conventions. His main aim was in line with Paul Grice's project to base the (elusive) notion of 'meaning' on beliefs, desires, and intentions of the agents of a conversation. As suggested in section 2 of this Introduction, signaling games have been studied extensively by economists to investigate, among others, under which circumstances a message credibly conveys information about the world. This research does not have a big impact yet on linguistics. In the first contribution to this book, **Robert Stalnaker** seeks to close that gap, by showing the analogy between Grice's philosophical analysis of meaning and the more recent game theoretical analysis of credible information exchange.

In the second chapter, **Prashant Parikh** introduces his games of partial information and argues that they extend signaling games. He shows how some pragmatic phenomena can be accounted for within his framework, and points out that game theory might be the appropriate tool to account for probabilistic communication. In the latter part of his paper, Parikh argues that the utterance situation s is not only important to contribute the game model required to calculate the semantic meaning of an utterance, but also to determine which solution concept is appropriate to use. He suggests that this can be accounted for in terms of (a sequence of) higher order games.

The main aim of this book is to show that game theory might shed new light on the study of language, mainly because it suggests that a very formal analysis of language use is within reach that takes a broader conception of language use than is standard in pragmatic analyses. However, by making use of game theoretical analyses, one also takes over its assumptions.

Nicholas Allott's paper contains a critical discussion of game theoretical analyses of communication. Because Prashant Parikh's analysis is the oldest and arguably best worked-out analysis of this sort, he naturally concentrates his discussion on this. Allott argues any analysis that makes use of standard game theory is based on some unmotivatedly strong assumptions, and suggests that some of these assumptions might be weakened by making use of some principles of Sperber and Wilson's (1986) Theory of Relevance.

Perhaps the main problem of game theoretical analysis of communication is the fact that such analyses typically predict that communication games have multiple equilibria, and that it is not a priori clear which one of those the conversational partners should, or will, coordinate on. A natural suggestion – also made by Prashant Parikh– is that of the various equilibria, agents typically converge to the Pareto optimal one, the equilibrium that gives to both participants the highest payoff. Natural as this proposal might seem, Sally (2003) has pointed out that in many game theoretical situations this is not the outcome we actually observe in case the preferences of the agents are not fully aligned. In those cases, avoidance of risk plays an important role as well. Following Sally's observations, **Robert van Rooij** and **Merlijn Sevenster** discuss the importance of risk for the use of expressions with an intended non-literal interpretation, or with an underspecified meaning.

The chapter by **Nicholas Asher** and **Madison Williams** investigates the rational basis for the computation of pragmatic interpretation from semantic content. They argue that an analysis of pragmatic inference in terms of Lewisian coordination games is insufficient because that model lacks a principled account of equilibrium selection. To overcome this problem, they develop a dynamic version of Bacharach's (1993) Variable Frame Theory, which in turn builds on Schelling's (1960) notion of focal points. The compositional interpretation of an utterance, together with the mutual world knowledge, defines a starting point in a game dynamics, which in turn converges on the pragmatic interpretation of the utterance. This approach is motivated and illustrated with several default inference patterns from Asher and Lascarides' (2003) Segmented Discourse Representation Theory.

Anton Benz's chapter explains the possibility of partial and mention-some answers in the context of two-person games. Starting out with Groenendijk and Stokhof's(1984) semantic approach he argues that their occurrence can be explained if we assume that they are embedded into contextually given decision problems. This builds on work by Merin (1999b) and especially van Rooij (2003b). He shows that intuitive judgments about the appropriateness of partial and mention–some answers are in accordance with the assumption that interlocutors are Bayesian utility maximizers. In the second part of his paper, he proves that explanations that are based on purely decision-theoretically defined measures of relevance cannot avoid picking out misleading answers.

The chapter by **Kris de Jaegher** shows that the grounding strategies of interlocutors can be characterized as evolutionarily stable equilibria in variants of the so-called electronic mail game (Rubinstein 1989). In conversation, it is not only necessary to achieve common knowledge about the meaning of utterances but also about the fact that some information has been communicated. The strategies employed by the interlocutors to achieve this goal are called their grounding strategies. Kris de Jaegher shows that separating equilibria in an electronic mail have a natural interpretation as grounding strategies. He shows especially that Traum's (1994) grounding acts are among the evolutionarily stable equilibria.

The chapter by **Jacob Glazer** and **Ariel Rubinstein** studies the rules of pragmatics in the context of a debate between two parties aiming to persuade a listener to adopt one of two opposing positions. The listener's optimal conclusion depends on the state of the world initially known only to the two parties. The parties argue sequentially. Arguing entails providing some hard evidence. A persuasion rule determines the conclusion that the listener will draw from the arguments made. A state of the world and a persuasion rule determine a zero-sum game played by the two debaters. The outcome of the game is the conclusion drawn by the listener, which might be right or wrong. The paper imposes a constraint on the amount of information that the listener can absorb and characterizes the persuasion rules that minimize the probability that the listener reaches the wrong conclusion. It is demonstrated that this optimization problem is affected by the language in which the persuasion rules are defined.

The last chapter in this volume, by **Tom Lenaerts** and **Bart de Vylder**, is of a somewhat different nature than the others. It concentrates not so much on the effects of our beliefs and preferences on what is communicated in an actual conversation, but rather on how a conventional language can emerge in which expressions have a meaning shared among a group of autonomous agents. It is the only contribution in this volume that makes use of the tools of evolutionary game theory. This paper discusses the effect of a particular model of language learning on the evolution of a conventional communication system. We feel that this chapter is especially suited to this volume, because – and this in contrast to almost all other analyses of the evolution of language that give great weight to language learning – language learning in this model is not supposed to be passive, and only used by children, but rather active, where the learner's language use also plays an important role.

Notes

1. The standard statistical relevance of a proposition E for a hypothesis H is defined by $R(H, E) = P(H/E) - P(H)$. The standard statistical relevance and Good's relevance are identical with respect to all properties that we use in this introduction, especially, it is $R(H, E) = -R(\overline{H}, E)$.
2. We can look at $\log(P^+(H)/P^+(\overline{H}))$ as a (possibly negative) measure for our inclination to favor H over $\overline{H}$; hence $r_H(E)$ tells us how the strength of this inclination is *updated*. This is an advantage of $r_H(E)$ over the standard statistical notion of relevance $P(H/E) - P(H)$.
3. See also Parikh's contribution to this volume.
4. Parikh (Parikh 1991, Parikh 2001) studies what he calls *Games of Partial Information* and claims in his contribution to this volume that they are more general than the signaling games as studied in economics and biology.
5. Or, more generally, as a set of subsets of T.
6. If hearers use such an interpretation rule, speakers have no reason anymore to be vague. But, of course, vagueness can still have positive pay-off when one's audience is unsure about your preferences.
7. See van Rooij and Schulz (2004) for more discussion.
8. In the standard model of EGT, populations are – simplifyingly – thought of as infinite and continuous, so there are no minimal units.
9. A *trajectory* is the path of development of an evolving entity.
10. One might argue that the strategies of a language user in these two roles are not independent. If this correlation is deemed to be important, the whole scenario has to be formalized as a symmetric game.
11. A similar point can be made with regard to the prisoners' dilemma, where the unique NE, general defection, is also the unique ESS, both in the symmetric and in the asymmetric conception.
12. A *Markov process* is a stochastic process where the system is always in one of finitely many states, and where the probability of the possible future behaviors of the system only depends on its current state, not on its history.
13. This idea was first developed in Kandori et al. (1993) and Young (1993). Fairly accessible introductions to the theory of stochastic evolution are given in Vega-Redondo (1996) and Young (1998).
14. Provided the population is sufficiently large, that is. Very small populations may display a weird dynamic behavior, but we skip over this side aspect here.
15. The system of difference equations used in the experiment is

$$\begin{aligned}\frac{\Delta x_i}{\Delta t} &= x_i((A\mathbf{y})_i - \langle \mathbf{x} \times A\mathbf{y}\rangle) + \sum_j \frac{Z_{ji} - Z_{ij}}{n} \\ \frac{\Delta y_i}{\Delta t} &= y_i((B\mathbf{x})_i - \langle \mathbf{y} \times B\mathbf{x}\rangle) + \sum_j \frac{Z_{ji} - Z_{ij}}{n}\end{aligned}$$

where $\mathbf{x}, \mathbf{y}$ are the vectors of the relative frequencies of the speaker strategies and hearer strategies, and A and B are the payoff matrices of speakers and hearers respectively. For each pair of strategies i and j belonging to the same player, Z_{ij} gives the number of individuals that mutate from i to j. Z_{ij} is a random variable which is distributed according to the binomial distribution $b(p_{ij}, \lfloor x_i n \rfloor)$ (or $b(p_{ij}, \lfloor y_i n \rfloor)$ respectively), where p_{ij} is the probability that an arbitrary individual of type i mutates to type j within one time unit, and n is the size of the population. We assumed that both populations have the same size.

16. Generally, if (s_i, h_j) form a SNE, the hearer has an incentive to deviate from it as soon as the speaker chooses a mixed strategy x such that for some $k \neq j, \sum_{i'} x_{i'} u_h(s_{i'}, h_k) > \sum_{i'} x_{i'} u_h(s_{i'}, h_j)$. The minimal amount of mutants needed to drive the hearer out of the equilibrium would be the minimal value of $1 - x_i$ for any mixed strategy x with this property. (The same applies *ceteris paribus* to mutations on the hearer side.)

Bibliography

Anscombre, J. C. and O. Ducrot (1983). *L'Argumentation dans la Langue*. Mardaga, Brussels.

Asher, N. and A. Lascarides (2003). *Logics of Conversation*. Cambridge University Press, Cambridge (UK).

Bacharach, M. (1993). Variable universe games. In K. Binmore, A. Kirman, and P. Tani, eds., *Frontiers of Game Theory*. MIT Press, Cambridge, MA.

Crawford, V. and J. Sobel (1982). Strategic information transmission. *Econometrica*, **50**, 1431–1451.

Ducrot, O. (1973). *La preuve et le dire*. Mame, Paris.

Ellison, G. (2000). Basins of attraction, long run equilibria, and the speed of step-by-step evolution. *Review of Economic Studies*, **67**(1), 17–45.

Farrell, J. (1988). Communication, coordination and Nash equilibrium. *Economic Letters*, **27**, 209–214.

Farrell, J. (1993). Meaning and credibility in cheap-talk games. *Games and Economic Behavior*, **5**, 514–531.

Fauconnier, G. (1975). Pragmatic scales and logical structure. *Linguistic Inquiry*, **6**, 353–375.

Frege, G. (1918). Der Gedanke: eine logische Untersuchung. *Beitrage zur Philosophie des deutschen Idealismus*, **1**, 58–77.

Gibson, R. (1992). *A Primer in Game Theory*. Harvester Wheatsheaf, Hertfordshire.

Good, I. (1950). *Probability and the Weighing of Evidence*. Griffin, London.

Grice, H. P. (1967). Logic and conversation. In *William James Lectures*. Harvard University. Reprinted in *Studies in the Way of Words*, 1989, Harvard University Press, Cambridge, MA.

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

Hirschberg, J. (1985). *A Theory of Scalar Implicatures*. Ph.D. thesis, University of Pennsylvania.

Horn, L. (1991). Given as new: When redundant affirmation isn't. *Journal of Pragmatics*, **15**, 313–336.

Horn, L. (1993). Economy and redundancy in a dualistic model of natural language. In S. Shore and M. Vilkuna, eds., *1993 Yearbook of the Linguistic Association of Finland*, pp. 33–72. SKY.

de Jaegher, K. (2003). A game-theoretical rationale for vagueness. *Linguistics and Philosophy*, **26**, 637–659.

Jäger, G. (2004). Evolutionary Game Theory and typology: a case study. Manuscript, University of Potsdam and Stanford University.

Kandori, M., G. Mailath, and R. Rob (1993). Learning, mutation, and long-run equilibria in games. *Econometrica*, **61**, 29–56.

Kreps, D. and R. Wilson (1982). Sequential equilibrium. *Econometrica*, **50**, 863–894.

Labov, W. (1972). *Sociolinguistic Patterns*. University of Pennsylvania Press, Philadelphia.

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

Lipman, B. (2003). Language and economics. In M. Basili, N. Dimitri, and I.Gilboa, eds., *Cognitive Processes and Rationality in Economics*. Routledge, London.

Lipman, B. and D. Seppi (1995). Robust inference in communication games with partial provability. *Journal of Economic Theory*, **66**, 370–405.

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

Merin, A. (1999a). Die relevance der relevance: Fallstudie zur formalen semantik der englischen konjuktion *but*. Habilitationschrift, University Stuttgart.

Merin, A. (1999b). Information, relevance, and social decision making: Some principles and results of decision-theoretic semantics. In L. Moss, J. Ginzburg, and M. de Rijke, eds., *Logic, Language, and Information*, volume 2, pp. 179–221. CSLI Publications, Stanford.

von Neumann, J. and O. Morgenstern (1944). *The Theory of Games and Economic Behavior*. Princeton University Press, Princeton.

Nowak, M. A., N. L. Komarova, and P. Niyogi (2002). Computational and evolutionary aspects of language. *Nature*, **417**, 611–617.

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

Parikh, P. (1992). A game-theoretical account of implicature. In Y. Vardi, ed., *Theoretical Aspects of Rationality and Knowledge*. TARK IV, Monterey, California.

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

Parikh, R. (1994). Vagueness and utility: The semantics of common nouns. *Linguistics and Philosophy*, **17**, 521–535.

Pratt, J., H. Raiffa, and R. Schlaifer (1995). *Introduction to Statistical Decision Theory*. The MIT Press, Cambridge, MA.

Rabin, M. (1990). Communication between rational agents. *Journal of Economic Theory*, **51**, 144–170.

van Rooij, R. (2003a). Being polite is a handicap: Towards a game theoretical analysis of polite linguistic behavior. In M. Tennenholtz, ed., *Proceedings of the 9th conference on Theoretical Aspects of Rationality and Knowledge*. ACM Press, New York.

van Rooij, R. (2003b). Questioning to resolve decision problems. *Linguistics and Philosophy*, **26**, 727–763.

van Rooij, R. (2004). Signalling games select Horn strategies. *Linguistics and Philosophy*, **27**, 493–527.

van Rooij, R. and K. Schulz (2004). Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information*, (13), 491–519.

Rubinstein, A. (1989). The electronic mail game: strategic behavior under 'almost common knowledge,'. *American Economic Review*, **79**, 385–391.

Sally, D. (2003). Risky speech: behavioral game theory and pragmatics. *Journal of Pragmatics*, **35**, 12231245.

Schelling, T. (1960). *The Strategy of Conflict*. Harvard University Press.

Selten, R. (1980). A note on evolutionarily stable strategies in asymmetric animal conflicts. *Journal of Theoretical Biology*, **84**, 93–101.

Sperber, D. and D. Wilson (1986). *Relevance. Communication and Cognition*. Basil Blackwell, Oxford.

Taylor, P. and L. Jonker (1978). Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences*, **40**, 145–156.

Traum, D. R. (1994). *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester.

Vega-Redondo, F. (1996). *Evolution, Games, and Economic Behaviour*. Oxford University Press, Oxford.

Young, H. P. (1993). The evolution of conventions. *Econometrica*, **61**, 57–84.

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

Zahavi, A. (1975). Mate selection — a selection for a handicap. *Journal of Theoretical Biology*, **53**, 205–213.

Zahavi, A. and A. Zahavi (1997). *The Handicap Principle. A missing piece of Darwin's puzzle*. Oxford University Press, Oxford.