Signaling under uncertainty
Brochhagen, T.S.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
Chapter 2
Signaling Games: Analysis and Interpretation

“I don’t know what you mean by ‘glory’,” Alice said. Humpty Dumpty smiled contemptuously. “Of course you don’t – till I tell you. I meant ‘there’s a nice knock-down argument for you!’”
“But ‘glory’ doesn’t mean ‘a nice knock-down argument’,” Alice objected. “When I use a word,” Humpty Dumpty said, in rather a scornful tone, “it means just what I choose it to mean – neither more nor less.”

Lewis Carroll, *Through the Looking-Glass*

Where our goal is to analyze the conditions that may give rise to linguistic properties in interaction, we first need to specify how the choices that make up such interactions are made. Specifying linguistic choice, in turn, requires the specification of interlocutors’ communicative goals, the context of interaction, and other aspects that may influence how these goals are reached. For instance, interlocutors’ beliefs about each other’s linguistic behavior. Game theory gives us the means to make these notions and their interplay precise.

The fundamentals of game theory were laid out in von Neumann and Morgenstern’s (1944) *Theory of Games and Economic Behavior*. In it, a game is understood as any interaction between agents for which all possible actions and their joint outcome can be specified. A straightforward case that satisfies these conditions is an interaction with simple and overtly acknowledged rules, actions, and goals, such as a game of rock-paper-scissors. Table 2.1 shows this game in normal form, which is read as follows. The possible moves of one player, call her player one, are represented by the table’s rows. The possible moves of the other player, call her player two, are given by the table’s columns. According to this specification each player can perform one of three actions: *rock*, *paper* or *scissors*. The rules of the game dictate that both players should reveal their choices simultaneously. The outcome of the combination of their choices is described by the
Chapter 2. Signaling Games: Analysis and Interpretation

Table 2.1: Normal-form representation of a game of rock-paper-scissors.

<table>
<thead>
<tr>
<th>Player 1</th>
<th>rock</th>
<th>paper</th>
<th>scissors</th>
</tr>
</thead>
<tbody>
<tr>
<td>rock</td>
<td>(0, 0)</td>
<td>(−1, 1)</td>
<td>(1, −1)</td>
</tr>
<tr>
<td>paper</td>
<td>(1, −1)</td>
<td>(0, 0)</td>
<td>(−1, 1)</td>
</tr>
<tr>
<td>scissors</td>
<td>(−1, 1)</td>
<td>(1, −1)</td>
<td>(0, 0)</td>
</tr>
</tbody>
</table>

Table’s cells, where a numeric value of 1 is attached to winning, −1 to losing, and 0 to a draw. In Table 2.1 the payoff of player one is given by the first number in a cell’s bracket and that of player two is given by the second number. In this way, we know that the pair of actions \( \langle \text{rock, rock} \rangle \) gives a payoff of \( (0, 0) \), a draw, whereas \( \langle \text{rock, scissors} \rangle \) is a win for player one and a loss for player two.

Even for this seemingly mundane game there are some interesting questions a game theorist may want to answer. For instance, one may ask what action a player should take given the information she has about her opponent’s behavior. If it is known, for example, that the opponent plays scissors half of the time. Another question one may ask is how an action policy should change over time. For example, after witnessing the opponent play rock 10 times in a row.

The broad conception of a game put forth by von Neumann and Morgenstern allows us to ask similar questions about other kinds of interactions. Indeed, game theory has found applications in a multitude of fields, ranging from economics, political science and biology to computer science and linguistics. In this chapter we discuss some of the questions that the conception of language (use) as a game can inform us about, as well as how we may go about answering them.

Section 2.1 doubles as an introduction to fundamental game-theoretic notions as well as to Lewis’ (1969) signaling games. Section 2.2 builds on these notions to characterize linguistic outcomes, using classic static solution concepts such as that of a Nash equilibrium. The shortcomings of static equilibrium analysis are discussed in detail in Section 2.3, and contrasted with dynamic analysis. This discussion motivates the kind of analysis we conduct throughout this investigation. Section 2.4 introduces a family of models of rational language use, which we employ in following chapters to characterize linguistic behavior. In particular, to characterize pragmatic inference. Final remarks on our main assumptions are given in Section 2.5.

### 2.1 Signaling Games

In his seminal work on conventions, Lewis (1969) laid out the backbone on which much of modern game-theoretic analysis of communication rests (see Skyrms 2010 for an overview). According to Lewis communication can conceived as a
2.1. Signaling Games

strategic endeavor between interlocutors. Central to his analysis are signaling games: formal characterizations of information transfer mediated by messages sent from speakers to addressees. The classical setup considers two agents: a sender and a receiver. The sender has some information that she wishes to convey to the receiver. As the receiver has no direct access to the sender’s information state, the sender has to resort to the use of messages. Upon reception of a message the receiver’s task is to act upon it.

Signaling games can characterize a variety of communicative situations. For instance, animal alarm calls. In the signaling literature, vervet monkeys are particularly famous for their alarm calls (e.g., Seyfarth et al. 1980, Cheney and Seyfarth 1990, Skyrms 2010:§2, Price et al. 2015; see Zuberbühler 2009 for an overview on animal alarm calls). A vervet alarm call depends on the type of predator observed. This allows receivers of the alarm call to take appropriate evasive action. For example, the alarm call for an aerial predator may effect the act of hiding in bushes. One used for terrestrial predators may instead be answered by hiding in trees. By analogy, in the case of human communication, the appropriate act to a request such as Pass me the salt would be to pass the salt (if the addressee is able and willing to do so). The particular kind of receiver acts we will focus on in this investigation are interpretations. That is, we assume the receiver’s act to be to interpret the message she receives as a particular state. This simplifies our notation a little, as we need not focus on sender states and receiver acts, but rather on a single set of states relevant to the context of interaction. More precisely, a classical signaling game considers a set of sender states $S$, a set of messages $M$, and a set of receiver acts $A$. An interpretation game is one where $A = S$.

The strategic aspect of interactions in signaling games lies in the choices made by each agent and the joint outcome they wish to effect. What message the sender sends for which state hinges on the receiver’s (expected) interpretation of the message. Conversely, the receiver’s interpretation hinges on the way in which messages are (expected to be) used by the sender. If interlocutors can coordinate in such a way that messages sent in a state are interpreted as conveying that state then information is transferred faithfully.

**Strategies.** More formally, a signaling game is a sequential two player game. In contrast to a game of rock-paper-scissors, this means that choices are not simultaneous. Instead, sender choices are contingent on states (they are in) and receiver choices are contingent on messages (they receive). In other words, a receiver’s interpretation follows a sender’s choice. A player’s complete contingency plan of which message/state to send/infer when is called a strategy. If these choices are deterministic, a sender strategy is a mapping from states to messages, $\sigma: S \rightarrow M$, and a receiver strategy is one from messages to states, $\rho: M \rightarrow S$. Such deterministic strategies are called pure in game-theoretic parlance.
We will also want to consider strategies that are not pure but instead capture probabilistic behavior. A probabilistic sender strategy is a mapping from states to probability distributions over messages, $\sigma : S \rightarrow \Delta(M)$. A probabilistic receiver strategy is one from messages to distributions over states, $\rho : M \rightarrow \Delta(S)$. We will often denote the probability of a sender choosing message $m$ given state $s$ under behavioral strategy $\sigma$ as $\sigma(m \mid s)$. Analogously, $\rho(s \mid m)$ is the probability of the receiver interpreting message $m$ as state $s$ under $\rho$.

At first sight, it may seem strange for interlocutors to adopt a probabilistic strategy over a pure one. After all, while there are situations in which players gain from unpredictability, such as when playing rock-paper-scissors, less predictability in communication might work against mutual understanding. Mutual understanding improves if the receiver can reliably infer what the sender wishes to convey and the better the sender can predict the receiver’s interpretative behavior. There are two general reasons why considering probabilistic strategies is nevertheless desirable. First, communication often involves a substantial amount of uncertainty, meaning that neither sender nor receiver can be certain about each other’s behavior. Probabilistic strategies can accordingly be conceptualized as resulting from conjectures about interlocutor behavior (Franke 2009:§1.2.3). Second, as made precise in Section 2.4, probabilities are convenient to represent an agent’s occasional deviation from rational behavior, failure to recognize slight differences between the outcome of different choices, other consequences of bounded rationality, or imperfect perception of the environment (see Chapter 6).

To make matters more concrete, consider a signaling game with two states, $s_1$ and $s_2$, and two messages, $m_1$ and $m_2$. In this game there are four pure sender strategies and four pure receiver strategies, listed in Table 2.2. The sequential nature of the game and the interdependence of sender and receiver strategies is illustrated in Figure 2.1.
Preferences and (expected) utility. Intuitively, some of the strategies listed in Table 2.2 are less suited for communication than others. For instance, a sender adopting strategy $\sigma_1$ will always send message $m_1$, irrespective of the state she is in. Conversely, a receiver using $\rho_1$ will interpret any message she receives as state $s_1$. By contrast, the strategy combinations $\langle \sigma_2, \rho_2 \rangle$ and $\langle \sigma_4, \rho_4 \rangle$ ensure that the state that the sender is in is also the one inferred by the receiver.

That sender and receiver care about matters such as successful information transfer is captured by the notion of utility, a subjective measure of an agent’s preference over outcomes of a game. More precisely, utility is a function from outcomes to real numbers: $U_{\sigma, \rho} : S \times M \times S \to \mathbb{R}$.

Beyond having a preference about which state is inferred when, interlocutors can also have preferences about how matters are communicated. A sender may prefer a polite but less clear expression over one that is more explicit; have particular stylistic predispositions; prefer shorter over longer expressions; or have preferences over matters such as the relative cognitive load of the retrieval of expressions. A receiver may have other, possibly opposed, preferences over messages.

Following the distinction between what is communicated and how it is communicated, signaling games are often distinguished depending on whether players have differential preferences over messages. Signaling games in which one message is as good as any other to all players are known as cheap talk, so called because no message carries cost (or all are equally costly). By contrast, if preferences over messages are relevant, message cost is represented as a small but non-negligible amount deducted from an interlocutor’s preference for the communicated state when the message is used. In other words, message cost is inverse to message preference but nominal. That is, small relative to preferences over what is communicated when (Blume et al. 1993, Benz and van Rooij 2007).

Even if signaling is free of cost, communicative preferences may be opposed. A sender could, for instance, prefer the receiver to always infer state $s_1$, irrespective of whether this is the actual state. The receiver might instead prefer to know the actual sender state. This situation could represent the preferences of an applicant who wants her interviewer to think she is qualified for a position even when she is not; that of a predator mimicking a harmless species to lure in prey; or that of prey mimicking a noxious species to avoid predation. In cases where players’ preferences are orthogonal to each other communication is not a cooperative affair. If, as standardly assumed, players’ preferences are common knowledge, a player’s best interest would then be to detect deceitful behavior and turn it to their advantage. Whether messages are credible and information transfer is possible at all will then depend on how aligned preferences are, what each message is (believed) to mean, what other messages are at an agent’s disposition, as well as other aspects of the interaction, such as message cost. In this investigation, we will instead be concerned with cooperative communication, meaning that sender and receiver strive for mutual understanding. In signaling games this is reflected by
utility functions that assign higher utility to successful information transfer than to misunderstanding, while leaving room for potentially diverging preferences over messages.

Taking stock, outcomes in signaling games are triples of a sender state $s$, a sent message $m$, and receiver interpretation $s'$. Letting $c_{\sigma,\rho}(\cdot)$ codify the subjective cost assigned to a message by sender $\sigma$ or receiver $\rho$, their utility can be defined as:

$$U_{\sigma}(s, m, s') = \delta(s, s') - c_{\sigma}(m); \quad (2.1)$$
$$U_{\rho}(s, m, s') = \delta(s, s') - c_{\rho}(m), \quad (2.2)$$

where

$$\delta(s, s') = \begin{cases} 1 & \text{if } s = s' \\ 0 & \text{otherwise} \end{cases} \quad (2.3)$$

In Gricean terms, $\delta(\cdot)$, as defined in (2.3), codifies cooperativity. In line with Grice’s (1975) postulated maxims of rational language use, it moreover follows from rational choice as utility maximization that a sender will convey matters, e.g., as clearly but succinctly as possible while avoiding false or misleading state-
ments (as long as individuals actually have such preferences). This game-theoretic rendering can accordingly capture the fundamental insights that underlie the Gricean program, but also allows us to go beyond it by enabling for the consideration of non-cooperative situations, as well as more differentiated preferences. We return to this issue and pragmatic inference more generally in Section 2.4.

If talk is cheap, preferences over outcomes in a cooperative interpretation game reduce to pairings of a sender state $s$ and a receiver’s interpretation $s'$. In a cooperative game these preferences are equal for both players. For a signaling game with two states they can be summarized by the matrix in Table 2.3.

Utility captures the quantitative preference of an agent for a single outcome. Expected utility takes this a step further and gives the weighted mean utility that interlocutors can expect when interacting. For finite $S$ and two players, $\sigma$ and $\rho$,
2.2 Static Concepts and Optimal Solutions

With these notions at hand we can return to our 2-states/messages signaling game and quantitatively compare strategy pairings. For expository ease, let us assume that talk is cheap, that each state is equally probable, $P^*(s_1) = 1/2 = P^*(s_2)$, and focus only on pure strategies. The expected utilities of this game’s 16 possible pure strategy combinations are given in Table 2.4.

Inspecting how well any two strategy pairings of this game fare reveals two things. First, many pairings leave no room for improvement. For instance, should a receiver be confronted with a sender that signals according to $\sigma_1$ then she can do no better than 0.5 under any strategy. To see this, inspect the row of $\sigma_1$ in Table 2.4. Put differently, a receiver following $\rho_3$, for example, has no incentive to change her interpretative behavior when interacting with $\sigma_1$. There is no alternative strategy that would yield better results. Second, the strategy combinations $\langle \sigma_2, \rho_2 \rangle$ and $\langle \sigma_4, \rho_4 \rangle$ guarantee the highest expected utility of 1.

---

Table 2.4: EU($\sigma_i, \rho_j$) in a cheap talk 2-states/messages signaling game.

<table>
<thead>
<tr>
<th>$\rho_1$</th>
<th>$\rho_2$</th>
<th>$\rho_3$</th>
<th>$\rho_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma_1$</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>$\sigma_2$</td>
<td>.5</td>
<td>1</td>
<td>.5</td>
</tr>
<tr>
<td>$\sigma_3$</td>
<td>.5</td>
<td>.5</td>
<td>.5</td>
</tr>
<tr>
<td>$\sigma_4$</td>
<td>.5</td>
<td>0</td>
<td>.5</td>
</tr>
</tbody>
</table>

---

expected utility is defined as:

$$EU_\sigma(\sigma, \rho) = \sum_s P^*(s) \sum_m \sigma(m|s) \sum_{s'} \rho(s'|m) U_\sigma(s, m, s');$$

(2.4)

$$EU_\rho(\sigma, \rho) = \sum_s P^*(s) \sum_m \sigma(m|s) \sum_{s'} \rho(s'|m) U_\rho(s, m, s'),$$

(2.5)

where $P^*(s)$ is the probability of state $s$. How $P^*$ is to be interpreted in linguistic terms has been subject to some discussion (Allott 2006, Franke 2013) and will become relevant in later chapters. To better explain what is at stake we defer this issue to Section 3.2.2. We may until then think of $P^*$ as an abstract exogenous determinant for what state the speaker is in, often referred to as nature.
That is to say that either pairing guarantees that states are always communicated successfully. In contrast to the first case, there is not only nothing to be gained from unilaterally adopting a different strategy but doing so would always be detrimental.

Both of these properties are central in game theory. The former, when there is no incentive to unilaterally switch to a different strategy, is called a \( \text{(weak)} \) Nash equilibrium. The latter, when no incentive for unilaterally switching exists and, moreover, doing so would always be disadvantageous, is called a \( \text{strict Nash equilibrium} \).

More formally, let \( s_i \) denote player \( i \)'s strategy and \( s_{-i} \) the strategies of all players except \( i \). Then \( \langle s^*_i, s^*_{-i} \rangle \) is a weak or, respectively, strict Nash equilibrium should there be no alternative strategy \( s_i \) for any player \( i \) such that

\[
EU_i(s^*_i, s^*_{-i}) \leq EU_i(s_i, s^*_{-i}); \quad \text{(2.6)}
\]

or

\[
EU_i(s^*_i, s^*_{-i}) < EU_i(s_i, s^*_{-i}). \quad \text{(2.7)}
\]

To reiterate in words, a Nash equilibrium is a collection of strategies in which no single player has an incentive to change her strategy provided everyone else conforms to theirs. In the context of signaling games, Lewis (1969) calls those Nash equilibria that are strict and lead players to associate each single state with a single and correct interpretation \( \text{signaling systems} \).

Central to a signaling system is that it is arbitrary. For the games Lewis focused on – cheap talk games with \( |M| = |S| = |A| \) and uniform \( P^* \) – there always exists at least one other equilibrium that results from a permutation of messages that is as optimal for information transfer as itself. Such is the relation between \( \langle \sigma_2, \rho_2 \rangle \) and \( \langle \sigma_4, \rho_4 \rangle \), depicted in Figure 2.2. Signaling systems consequently do not require messages to be meaningful in themselves. After all, messages can be used to signal completely different states and either signaling system is equally good for information transfer. Instead, what endows messages with meaning is their use in equilibrium because interlocutors behave “as if” they meant something. Moreover, although a signaling system is arbitrary, being a strict Nash equilibrium ensures that no individual would wish to unilaterally deviate from it.

As shown in Table 2.4, there are 2 signaling systems in a 2-states/messages game. More generally, in a \( N \)-states/messages signaling game, there are \( N! \) signaling systems.
2.2. Static Concepts and Optimal Solutions

Signaling systems intuitively correspond to desirable linguistic outcomes. However, the characterization of strategy combinations in terms of equilibria is silent on how they are reached. This begs the question whether they can be reached at all; and if so, under which conditions. One common conception of Nash equilibria is that they correspond to stable states when games are repeatedly played (Osborne 2004). Under this view, equilibrium analysis is agnostic about the process that leads players to adopt their behavior but predicts that, over time, players will reach equilibrium and thereafter remain in it. Players in equilibrium stay in equilibrium. Importantly, equilibrium analysis does not appeal to the rationality of players in reaching a particular outcome.

Populations and types. Nash equilibria describe optimal stable outcomes with respect to individual strategy pairings. To transfer this idea to a population of communicating agents we either have to restrict our attention to senders only meeting receivers and vice-versa, i.e., consider two distinct populations and their interaction, or alternatively endow agents with both a sender strategy and a receiver strategy. The latter is an arguably natural assumption for much of biological communication and is what we will assume throughout this investigation.

Irrespective of the nature of the population analysis employed, we call the units that distinguish members of a population types, \( \tau \in T \). Under a biological interpretation types can be identified with phenotypes and their expected utility with their fitness. This determines a type’s chance of survival and reproduction. Under a cultural interpretation a type corresponds to a particular linguistic behavior. That is, a sender and receiver strategy. As exemplified by monkey alarm calls (e.g., Seyfarth et al. 1980, Cheney and Seyfarth 1990, Skyrms 2010:§2), communicative success can be a determinant for survival and reproduction, closing the gap between communicative and biological fitness; being able to act upon an alarm call correctly can greatly enhance chances of survival. However, where the focus lies on human communication, my preferred interpretation is that agents themselves strive toward efficient and successful information transfer. They therefore occasionally adapt or revise their behavior to improve their communication with other members of the population (Benz et al. 2006b:§3.3, Skyrms 2010:55). In a dynamic setting a type’s higher fitness then translates to a higher chance that other agents will attempt to adopt/imitate this behavior (see Chapter 4 for details).

The transformation of a game to one in which all players draw from the same strategy pool is called its symmetrization (Wärneryd 1993, Cressman 2003, another cultural interpretation, put forth by Nowak et al. (2002), is that successful communication increases the chances of influencing the acquisition process of future generations. In the same fashion as my preferred interpretation, populations would come to reflect a higher proportion of successful past behaviors (assuming all types are equally likely to be acquired; see Chapter 4). Rather than this behavior being adopted by more agents horizontally, this would then be a consequence of the influence of their success on future generations.
Table 2.5: EU(τ_\text{i}, τ_\text{j}) in a symmetrized cheap talk 2-states/messages game.

Franke and Wagner (2014). While our illustrative 2-states/messages signaling game has 4 pure sender strategies and 4 pure receiver strategies, there are 16 types in its symmetrized counterpart. These result from pairing all possible sender and receiver strategies. Mnemonically labeling each type with an index of its sender and receiver strategy we have that \( τ_{11} = \langle σ_1, ρ_1 \rangle \) and that \( τ_{42} = \langle σ_4, ρ_2 \rangle \), for example. That is, \( τ_{11} \) sends and receives according to \( σ_1 \) and \( ρ_1 \) whereas \( τ_{42} \) follows \( σ_4 \) and \( ρ_2 \).

Assuming that agents are senders half of the time, the expected utility of \( τ_\text{i} \) interacting with \( τ_\text{j} \) is defined by:

\[
\text{EU}(τ_\text{i}, τ_\text{j}) = \frac{1}{2} \text{EU}_σ(τ_\text{i}, τ_\text{j}) + \frac{1}{2} \text{EU}_ρ(τ_\text{i}, τ_\text{j}).
\]

The expected utilities of the 16 types of the symmetrized 2-states/messages signaling game are given in Table 2.5.

Finally, fitness indicates how well a type communicates in a population. Letting \( x_\text{j} \) be a population vector with \( x_\text{j} \) corresponding to the proportion of \( τ_\text{j} \) in \( x \), the fitness of \( τ_\text{i} \) is defined as the average expected communicative success of this type given the type frequencies of the current population \( x \):

\[
f_\text{i} = \sum_j x_\text{j} \text{EU}(τ_\text{i}, τ_\text{j})
\]

Evolutionary stable strategies. Returning to the question about a population-level counterpart to a static solution concept such as that of a Nash equilibrium, a type’s stability is classically associated with the notion of invasibility (Maynard Smith and Price 1973). Intuitively, an evolutionary stable strategy (ESS)
is a strategy such that, if a population consist mostly of it, then this population is resistant against invader types changing the population’s composition. More simply put: a type is evolutionary stable if its population cannot be taken over by others. This concept, just as that of a Nash equilibrium, is static: it does not make explicit which processes may lead a different type to join or take over the population. Similarly, it does not appeal to the rationality of members of a population; nor their knowledge of other types behavior; nor to their knowledge of the structure of the game. Instead, types are compared solely based on their (expected) utility following the intuition that a strategy is an ESS if fitness-based selection is sufficient to drive out invaders.

The concept of an ESS is a refinement of that of a Nash equilibrium. Consider a population consisting solely of $\tau_{11}$. While $\langle \tau_{11}, \tau_{11} \rangle$ is a Nash equilibrium, such a population is tolerant toward any proportion of the 15 remaining types being present in it. As before, inspect the row of $\tau_{11}$ in Table 2.5 to see this. For instance, if an agent of type $\tau_{22}$ entered this population it would do as well as $\tau_{11}$-players. What is more, not only would $\tau_{22}$ not be driven out of the population, if more players of this type emerged then $\tau_{22}$ would do better than $\tau_{11}$. This would allow it to invade this initially monomorphic population. In short, Nash equilibria are too permissive to qualify as ESS. A possible solution would be to instead identify only strict Nash equilibria with stability at a population level. This move turns out to be too restrictive. Many games have mixed Nash equilibria that should count as stable (at the individual level; probabilistic strategies), but mixed Nash equilibria cannot be strict. In light of these considerations, Maynard Smith and Price (1973) propose that $\tau_i$ is an ESS iff

$$EU(\tau_i, \tau_i) \geq EU(\tau_j, \tau_i) \quad \text{for all alternative types } j; \quad (2.10)$$

$$EU(\tau_i, \tau_i) = EU(\tau_j, \tau_i) \rightarrow EU(\tau_i, \tau_j) > EU(\tau_j, \tau_j). \quad (2.11)$$

In words, $\tau_i$ is an ESS if, according to (2.10), it fares as least as well when interacting with itself than does any other type. Additionally, according to (2.11), should there be a type that fares as well against $\tau_i$ as $\tau_i$ against itself, then that type fares worse against its own type. The relationship between Nash equilibria and ESS can be summarized as follows (Nowak 2006):

strict Nash equilibrium $\Rightarrow$ ESS $\Rightarrow$ weak Nash equilibrium

A second common take on ESS is to make the degree to which an evolutionary outcome is impervious to invasion explicit, using an invasibility threshold $\bar{\epsilon}$. In this case, $\tau_i$ is an ESS iff for all $\epsilon < \bar{\epsilon}$ and all alternative types $j$:

$$\epsilon EU(\tau_i, \tau_j) + (1 - \epsilon)EU(\tau_i, \tau_i) > \epsilon EU(\tau_j, \tau_j) + (1 - \epsilon)EU(\tau_j, \tau_i). \quad (2.12)$$

That is, if the proportion of invaders is below threshold $\bar{\epsilon}$ then a population consisting of a proportion of $(1 - \epsilon)$ types $\tau_i$ cannot be taken over. The underlying intuition of this definition remains the same as that of conditions (2.10) and (2.11).
Chapter 2. Signaling Games: Analysis and Interpretation

2.3 On the Use and Limits of Static Equilibrium Analysis

Nash equilibria and ESS are informative about optimal outcomes at the individual and population level, respectively. The agnosticism of either type of static equilibrium analysis with respect to rationality and underlying processes is part of its appeal. Were it demonstrable that static predictions coincided with dynamic ones under fairly general conditions, our task would be restricted to that of finding a suitable model for a linguistic phenomenon under scrutiny and to then identify an appropriate static solution concept. There are, however, three major issues that static equilibrium methodology faces at either level. The first two are technical (Huttegger and Zollman 2013); the third is conceptual (Franke 2013).

Unclear predictions. First, as illustrated by the simple 2-states/messages case above, many signaling games have multiple equilibria. As mentioned earlier, Lewis’ (1969) setup in fact guarantees a multiplicity of strict Nash equilibria at the individual level. The existence of multiple equilibria percolates to the population level in these cases as well, as only these strict Nash equilibria are ESSs (Wärneryd 1993). Prima facie, static equilibrium analysis does not offer guidance as to what its predictions are when more than one solution exist. This leaves such an analysis at best incomplete.

Uncertain predictions. With Searcy and Nowicki (2005) one may nevertheless intuit that optimal signaling is expected to emerge in Lewisian signaling games because agents have a common interest in information transfer. Any equilibrium would do, and there might appear to be little mystery to the fact that one of them will emerge. If so, a dynamic analysis would then “only” inform us about the process that leads to the adoption of one equilibrium over the other (see the third issue discussed below on this matter). However, a second problem of static equilibrium analysis is that neither at the individual nor at the population level these outcomes necessarily obtain.

Two simple and well-studied processes that illustrate this issue are Roth-Erev reinforcement learning, at the individual level, and the replicator dynamic, at the population level. These processes are particularly relevant to this issue because both, in principle, promote high utility behavior. They may therefore be expected to favor the optimal outcomes predicted by static equilibrium analysis.

Roth-Erev reinforcement learning is a simple learning process whereby actions that were successful in the past are rendered more likely to be chosen in the future (Roth and Erev 1995, Erev and Roth 1998). In the context of signaling games this means that a successful interaction will lead to a stronger association of the conveyed state with the used message in the case of senders, and to a stronger...
2.3. On the Use and Limits of Static Equilibrium Analysis

Prior to any interaction, each agent’s associations are initialized with a weight. After an interaction relevant weights are then modified by a reinforcement value $r$. In each interaction the probability of choosing an action – to send a message or to interpret a message as a state – is proportional to its weight. Figure 2.3 illustrates this process. Initially, choice is driven by chance or whatever information fed the initial weight of a state-message or message-state association. However, over time, actions that were previously successful are more likely to be taken.

The replicator dynamic is one of the most famous population dynamics in evolutionary game theory. It models fitness-based selection, where the relative frequency of a type in a population increases with a gradient proportional to its average fitness in the population (Taylor and Jonker 1978, Hofbauer and Sigmund 2003; see Chapter 4 for details). This dynamic is popular and versatile because it can be derived from many abstract processes of biological and cultural transmission and selection (for overview and several derivations see Sandholm 2010), including conditional imitation (e.g., Helbing 1996, Schlag 1998) or reinforcement learning (e.g., Börgers and Sarin 1997, Beggs 2005).

Figure 2.4 shows the predictions of Roth-Erev reinforcement learning and the replicator dynamic for the cheap talk 2-states/messages signaling game. As showcased, neither dynamic guarantees the outcome predicted by its respective static equilibrium analysis. That is, not all dyads/populations converge to signaling systems. Instead, the proportion of suboptimal outcomes increases with the difference in frequency between the two states (see Catteeuw and Manderick 2014 for detailed analysis of reinforcement learning in signaling games, and Huttegger 2007 and Pawlowitsch 2008 on the replicator dynamic).

Many variants of these processes have been studied in connection to signaling games. For instance, reinforcement may not only be positive but punish actions that led to failure (e.g., Roth and Erev 1995). Alternatively, the stronger association of a state with a message may decrease this state’s association with other messages (e.g., Franke and Jäger 2011). The replicator dynamic can similarly be supplemented by perturbations in the form of type mutations at generational
Figure 2.4: Proportion of signaling systems established under Roth-Erev reinforcement learning and the replicator dynamic in a cheap talk 2-states/messages signaling game across state probabilities $P^* \in \Delta(S)$. Reinforcement results correspond, for each probability of $s_1$, to the mean outcome of 20000 independent games of 500 interactions each with initial weights set to 1 for all associations and $r$ equal to players’ utility. Dyads with an expected utility higher than .95 after 500 interactions were categorized as having established a signaling system. Replication results correspond to mean proportions of 20000 independent populations after 500 generations, each randomly initialized.

turn-overs; or by dropping the assumption that players are equally likely to encounter each other, instead correlating encounters with expected utility or an environment’s topological space (see, e.g., Skyrms 2010:§5). Some of these variations are more likely to converge to optimal outcomes than others. The crucial point, however, is that making these dynamics explicit is necessary to understand under which conditions an outcome can be expected to obtain. As summarized by Skyrms (2010:72): “The emergence of a signaling system is not always a moral certainty.”

**Explanatory force.** The third problem of static equilibrium analysis has to do with the quality of its explanatory potential. As argued by Franke (2013) appeals to optimality are unsatisfying if the conditions under which an outcome emerges are not made explicit, or if the assumptions underlying a solution concept lack justification. Put differently, even if Searcy and Nowicki’s (2005) intuition were correct, we would arguably not have learned much about the emergence and stability of a communication system. Instead, an explanatory analysis should be
able to answer questions such as how rational agents are required to be, what information they base their choices on, what adaptive mechanisms they (minimally) need to possess, and so on. By virtue of its procedural agnosticism, static analysis cannot answer these questions.

With these issues in mind, we should stress that static solution concepts are nevertheless useful to characterize outcomes with respect to measures of interest. For example, communicative success. Outcomes identified in this manner can then be critically compared to those borne out under particular processes, as well as to those that are evidenced empirically. Under this view, static equilibrium analysis is supplementary to dynamic analysis (Huttegger and Zollman 2013).

2.4 Individual-Level Behavior: Rational Language Use

Iterated interactions and population dynamics build on individual-level behavior. At this level dynamic alternatives to static solution concepts also exist, e.g., rationalizability (Bernheim 1984, Pearce 1984), Benz and van Rooij's (2007) Optimal Assertions model, the Rational Speech Act model (Frank and Goodman 2012, Goodman and Stuhlmüller 2013), and the iterated $X$-response family: iterated Best-, Cautious-, and Quantal-response (Jäger 2007b, Franke 2009, Franke and Jäger 2014). Common to these concepts is that they model players' choices in a single interaction as resulting from a reasoning process over beliefs about other players' behavior. This contrasts with the preceding exposition where individuals were assumed to follow a strategy without specifying why.

The idea that linguistic choice results from a process of mutual reasoning about rational language use brings us back to the Grice. However, there is an important difference between Grice's approach and the one pursued here, which we briefly touched upon earlier. According to Grice (1975; 1989) the process of reasoning that interlocutors engage in is guided by the mutual assumption that certain principles of (rational) language use are followed; e.g., to be succinct, orderly, and relevant (see Chapter 1). Most dynamic game-theoretic approaches to pragmatics also view mutual reasoning as the motor that leads interlocutors to their choices. However, they do not rely on the formulation of conversational principles to constrain and guide communication. Instead, these approaches ground linguistic choice in individuals' contextual beliefs and preferences. Under this view, the explanation of linguistic behavior is not in term of maxims but in terms of rational behavior according to such beliefs and preferences (Benz 2006, Benz and van Rooij 2007, Franke and Jäger 2016b:120f). In informal terms, in cooperative communication interlocutors do their best to ensure faithful information transfer according to their beliefs about others' linguistic behavior and their own communicative preferences. Not because they believe that a particular rule of conversation is being followed. In virtue of not being tied to explicit conversa-
tional rules, such an approach also allows for predictions in situations not covered by Grice’s principles, such as non-cooperative communication (e.g., De Jaegher and van Rooij 2014, Ahern and Clark 2017).

In recent years the idea that pragmatic reasoning can be captured by models that characterize linguistic inference as a product of explicit representations of mutual reasoning about beliefs and preferences has led to a diverse and growing literature. On a general level, this approach to pragmatics as rational language use encompasses game-theoretic approaches (e.g., Parikh 1991; 2000, Benz 2006, van Rooij and Sevenster 2006, Benz and van Rooij 2007, Jäger 2007b, Franke 2013, De Jaegher and van Rooij 2014, Franke and Jäger 2014) as well as Bayesian approaches (e.g., Frank and Goodman 2012, Goodman and Stuhlmüller 2013, Franke and Degen 2016, Bergen et al. 2016, Goodman and Frank 2016). Franke and Jäger (2016a:§3) identify five central properties common to these approaches: They are probabilistic, interactive, rationalistic, computational, and data oriented. The first four properties should not come as a surprise in light of the preceding discussion. The fifth refers to the fact that these models are usually not constrained to categorical predictions but allow for finer-grained quantitative predictions. This enables for a closer fit between theoretical predictions and actual communicative behavior evidenced, e.g., in experiments with human participants (Franke and Jäger 2016a:14). We will take advantage of this ability in Chapter 3. The remainder of this chapter incrementally introduces the common elements that make up models of rational language use.

2.4.1 Focal points and semantic meaning

For information transfer to take place in single interactions, the reasoning process that sender and receiver engage in should lead them to behave in a congruent fashion. That is, if things go smoothly, the receiver’s interpretation of a message should agree with the information state the sender is attempting to convey. However, only a belief in common rationality is too weak to ensure meaningful predictions in one-shot signaling games (see Franke 2009:§1.2 for details on this negative result). Intuitively, the problem is that many signaling games do not have a unique “obvious” solution that sender and receiver can reach independently, without prior communication and only through mutual reasoning. Put differently, these games lack a solution that is noteworthy relative to others in that it would allow all interlocutors to conclude the others’ reasoning process to be drawn to it. This issue was already touched upon from a different angle in Section 2.3, where we saw that Nash equilibria do not fulfill the requirement of uniqueness which could otherwise make them stick out relative to other strategies. To get meaningful inference without prior interaction off the ground the relation between the set of information states $S$ and the set of messages $M$ needs to be constrained.
2.4. Individual-Level Behavior: Rational Language Use

Figure 2.5: Illustration of a coordination problem without prior communication.

Focal points

Imagine that you and your partner are independently shown the four squares depicted in Figure 2.5 and that you have no means to communicate with each other. You are told that you will win a prize if both of you pick the same square. Which square would you choose? Similarly, imagine that you get lost in Amsterdam’s city center and cannot get in touch with your partner. Where would you attempt to meet?

At a fundamental level these problems are analogous to the coordination problem posed by signaling games: You and your partner care about performing congruent actions but in principle any behavior/square/location could be chosen. Crucially and differently from the games discussed so far, in the coordination problems sketched above some squares/locations will appear to be more obvious candidate solutions than others. This may seem particularly evident for the problem illustrated in Figure 2.5. It is therefore important to stress that, in terms of the problem’s setup, there is no reason to pick one square over the other. Coordination on a particular square is payoff irrelevant. You win the prize if you agree on any of them. However, things are different because you may believe that the third square draws you partner’s attention and the belief that she believes it draws yours as well is a good reason to choose it. Such beliefs need not but may well be partner specific. For instance, while Amsterdam’s Dam Square may be a more obvious location to rendezvous for residents, Amsterdam central station may be more salient to tourists.

Coordination problems such as these are what Schelling (1960) uses to illustrate the idea of focal points: solutions that are prominent, conspicuous or salient. In virtue of drawing reasoners’ attention such solutions can improve coordination in the absence of prior agreement and lack of utility-relative differentiation (see, e.g., Mehta et al. 1994 for behavioral experiments supporting this claim). As conceived by Schelling, focal points come into play to break a tie between strategies. They are consequently of particular relevance when the game itself lacks such a tie breaker. For our purposes, the question is then whether within communication there is a plausible constraint on the beliefs interlocutors’ may entertain about each other’s linguistic behavior. As argued by Franke (2009:§2.1.1), semantic meaning is a natural candidate to serve this purpose.
Semantic meaning

The idea that semantic meaning plays a role in constraining what messages are admissible in which information state, although not necessarily its conception as a focal point, has figured in many game-theoretic approaches to rational language use (e.g., Parikh 1992; 2000, Benz 2006, Jäger 2007b, Stalnaker 2006, Benz and van Rooij 2007) and bleeds into classic literature on message credibility in signaling games (e.g., Rabin 1990, Matthews et al. 1991, Farrell 1993). In informal terms, the problem of message credibility is that, prima facie, the receiver recognizing that the sender intends her to believe the meaning of a message does not need to imply that she will form this belief (Stalnaker 2006). A message is credible in case the receiver can reason the sender to send it only in case it is true. Put differently, message credibility concerns the question whether there are situations in which the sender has incentive to be untruthful. As mentioned earlier when discussing preferences and cooperation, this problem is pressing if there is at least some conflict of interest between interlocutors. If instead the game is (believed to be) one of pure coordination there is no reason to suspect foul play.

Message credibility is particularly important in the context of fully rational agents and solution concepts, particularly epistemic ones, for it answers the question whether reasoning stays within the boundaries of truth. As already fore-shadowed in Section 1.2, our analysis will not be concerned with these matters. Nevertheless, in line with this literature, semantic conventions and some degree of rationality will play a crucial role in providing the necessary fuel for meaningful pragmatic inference. In all generality, giving up the assumption of full rationality and allowing agents to occasionally err does however mean that we have to give up the claim that messages will always be used truthfully. In practice, the degrees of rationality we assume, while far from full rationality, will be sufficient to ensure that there is a strong tendency toward truthfulness in the games analyzed.

Returning to the conception of semantic meaning as a focal point, this view ascribes semantics a similar role to that of precedence and salience: it biases an agent’s decision procedure. As illustrated below, such a bias proves to be sufficiently restrictive to reign in the multiplicity of inferences otherwise obtained, while providing sufficient flexibility as to not necessarily determine the outcome of the reasoning process on which it is based. Different from Schelling’s (1960) conception of a focal point as a concept that applies at the final stage of deliberation, semantic meaning serves as a constraining point of departure for pragmatic reasoning. In other words, semantics provides the initial focal points that guide and constrain pragmatic inference.

In models of rational language use, semantic meaning is standardly represented by a lexicon $L$ which maps state-message pairs to the (Boolean) truth-value of the message in that state. As a model’s object, a lexicon represents the relevant semantic information required for pragmatic reasoning to get off the ground, relative to the phenomenon at hand. As illustrated shortly, the addition
of individuals’ beliefs and preferences to the model then yields the minimal set of relevant context features for pragmatic reasoning. A full specification of a game’s lexicon together with the sender’s and receiver’s cognitive make-up can accordingly be construed as a context model (Franke and Jäger 2014:§1) for a class of pragmatic inferences (e.g., scalar implicatures). A convenient way to represent lexica is by $(|S|,|M|)$-Boolean matrices.

### 2.4.2 Pragmatic reasoning

Models of rational language use have been applied to many linguistic phenomena, ranging from implicatures (e.g., Benz and van Rooij 2007, Jäger 2007b, Franke 2009, Frank and Goodman 2012, Franke and Jäger 2014, Bergen et al. 2016) and disambiguation (e.g., Parikh 2000) to the use of generics (Tessler and Goodman 2016), polite language use (Yoon et al. 2016) and prosodic emphasis (Bergen and Goodman 2015), to name a few. To better showcase commonalities and differences across models, we will focus on one prominent kind of pragmatic inference that has received substantive attention in the literature: scalar implicatures.

**Scalar implicatures**

Scalar implicatures are a particularly productive and well studied class of systematic pragmatic inferences (Horn 1984, Hirschberg 1985, Levinson 1983, Geurts 2010). Usually, the utterance of a sentence like *I own some of Johnny Cash’s albums* will be taken to mean that the speaker does not own all of them. This is because, if the speaker instead owned them all, she could have used the word *all* instead of *some* in her utterance, thereby making a more informative statement. Weak scalar expressions such as *some* are often semantically characterized as having literal meanings that are compatible with that of more informative relevant alternatives, like *all*. That is, if it were true that I own some of the albums, the literal meaning of *some* would not rule out that I own all of them. However, the use of a less informative expression when a more informative one could have been used can license a pragmatic inference that the stronger alternative does not hold. This rationale has been proposed to underlie the pragmatic use of many so-called scalar expressions. As exemplified by (2a) – (4a) and the defeasible scalar inferences they can give rise to in (2b) – (4b), English examples include numerals such as *five* and *six*, scalar adjectives like *warm* and *big*, as well as other quantifiers like *many*.

(2) a. I own some/many of Johnny Cash’s albums.
   b. $\sim$ I do not own all of Johnny Cash’s albums.

(3) a. I have five dogs.
   b. $\sim$ I do not have six/seven/... dogs.
Chapter 2. Signaling Games: Analysis and Interpretation

(4)  a. The soup is warm.
    b. \( \sim \) The soup is not hot.

Scalar implicatures, especially the inference from *some* to *some but not all*, have been studied extensively, both theoretically (e.g., Horn 1984, Sauerland 2004, van Rooij and Schulz 2004, Chierchia et al. 2012) as well as experimentally (e.g., Bott and Noveck 2004, Huang and Snedeker 2009, Grodner et al. 2010, Goodman and Stuhlmüller 2013, Degen and Tanenhaus 2015). While there is much dispute in this domain about many details, a clear majority endorses the view that a weak scalar item like *some* is underspecified to semantically mean *some and maybe all* and that the enrichment to *some but not all* is part of some regular process with roots in pragmatics (see Chapter 4 for further discussion and analysis).

A minimal game-theoretic rendering of the semantics relevant for scalar implicatures is as follows: Assume that there are two relevant world states \( S = \{ s_{\exists \neg \forall}, s_{\forall} \} \). In state \( s_{\exists \neg \forall} \) Chris owns some but not all of Johnny Cash’s albums while in \( s_{\forall} \) Chris owns them all. Additionally, assume that there are two relevant messages \( M = \{ m_{\text{some}}, m_{\text{all}} \} \), where \( m_{\text{some}} \) is short for a sentence like *Chris owns some of Johnny Cash’s albums* and \( m_{\text{all}} \) is short for the same sentence with *some* replaced by *all*. Lexica for this case would assign a Boolean truth-value, either 0 for false or 1 for true, to each state-message pair. The majority view of semantically underspecified *some* is captured by the following lexicon:

\[
L = \begin{bmatrix}
 s_{\exists \neg \forall} & m_{\text{some}} & m_{\text{all}} \\
 s_{\forall} & 1 & 0 \\
 s_{\forall} & 1 & 1
\end{bmatrix}
\]

In words, according to \( L \) the message \( m_{\text{some}} \) is true of both \( s_{\exists \neg \forall} \) and \( s_{\forall} \) whereas message \( m_{\text{all}} \) is true only of state \( s_{\forall} \).

**Rational language use**

What behavior would a rational user of \( L \) exhibit? A rational hearer would reason about the message she receives from the speaker, taking as a point of departure the semantics of \( L \). Intuitively and in analogy to the characterization of scalar implicatures given above, if a rational speaker means to convey \( s_{\forall} \) she should send \( m_{\text{all}} \). This message is semantically unequivocal, thereby increasing the chance of communicative success. In the case of \( s_{\exists \neg \forall} \), semantically, only message \( m_{\text{some}} \) is an option, but this message could signal either state. Nevertheless, if \( m_{\text{all}} \) is reasoned to signal \( s_{\forall} \), then the hearer can infer that \( m_{\text{some}} \) is to be associated with \( s_{\exists \neg \forall} \) not with \( s_{\forall} \). A rational speaker reasons in analogous fashion, coming to her behavior through reasoning about the hearer’s likely interpretation of a message. She will accordingly send \( m_{\text{some}} \) in \( s_{\exists \neg \forall} \) and \( m_{\text{all}} \) in \( s_{\forall} \).
2.4. Individual-Level Behavior: Rational Language Use

Models of rational language use such as the Optimal Assertions model (Benz 2006, Benz and van Rooij 2007), the Rational Speech Act model (Frank and Goodman 2012, Goodman and Stuhlmüller 2013) and Iterated Best-, Cautious- and Quantal-Response models (Jäger 2007b, Franke 2009, Franke and Jäger 2014) represent this reasoning procedure as a hierarchy over reasoning types. The bottom of the hierarchy, level 0, corresponds to literal signaling behavior. Literal language users do not reason about their interlocutors but simply produce/comprehend according to their preferences/expectations and the semantics provided by their lexica. Player $i$’s literal receiver and sender behavior are defined in (2.13) and (2.14).

$$\rho_0(s|m; pr^i; L) \propto L_{(s,m)} pr^i(s);$$  \hspace{1cm} (2.13)

$$\sigma_0(m|s; L) \propto L_{(s,m)} - c^\sigma(m),$$  \hspace{1cm} (2.14)

where $pr^i$ is player $i$’s prior over states, $pr^i \in \Delta(S)$. Such a prior represents the subjective relative saliency of states in a context. They are an individual’s expectations in a particular situation. We construe such expectations as based on any source of information beyond an expression’s literal meaning. Among others, a prior may draw from the context in which communication takes place, general expectations of language use, or perceptual information. In short, it represents condensed information of the association strength with which an interpretation comes to mind (Franke 2009:129ff). Whenever cost, priors, or lexica are assumed to be common – i.e., shared across individuals – we omit individual indices as well as their explicit codification as conditional parameters in choice probabilities.

To the best of my knowledge, in the case of scalar implicatures no strong evidence about differences in preferences between relevant message alternatives has been found, for example, between the choice of some and all; nor have particular prior biases been suggested. We may therefore assume that the scalar inference game is a cheap talk game with an uninformative common prior, $pr(s_u) = 1/2 = pr(s_{\exists \neg \forall})$. These assumptions are further motivated by the broader question whether some pragmatic inferences can be explained purely in terms of mutual reasoning. That is, whether certain inference patterns can be explained without appeal to state saliency or differential message preferences. This should not be taken to suggest that cost or priors should not play an explanatory role in pragmatics. Nor that an explanation that appeals to these factors is subordinate to one that does not. Rather, in the face of a lack of evidence to the contrary, an explanation that does not require a particular prior or cost-assignment is more parsimonious than one that does not.

Under these assumptions and with $L$ as above, definitions (2.13) and (2.14) give the following choice probabilities for literal behavior in the some-all context
model:

\[
\sigma_0(\cdot | \cdot) = \begin{cases} 
  m_{\text{some}} & 1 \\
  m_{\text{all}} & .5 
\end{cases} \quad \rho_0(\cdot | \cdot) = \begin{cases} 
  s_\exists \forall & .5 \\
  s_\forall & .5 
\end{cases}
\]

In words, a literal sender with no preferences over messages will be indifferent between sending \(m_{\text{some}}\) or \(m_{\text{all}}\) for \(s_\forall\), sending either with equal probability, and she will send \(m_{\text{some}}\) to convey \(s_\exists \forall\). Analogously, if their prior is flat, literal interpreters choose an arbitrary true interpretation for each message according to their lexicon.

Most models of rational pragmatic language use assume that naïve level-0 reasoners do not correspond to actual linguistic behavior. This is reflected by the fact that neither choice rule involves (an approximation of) utility maximization. Instead, (2.13) and (2.14) characterize completely unreflected literal language use.

Higher order reasoning types of level \(n + 1\) make their linguistic choices according to the expected behavior of a level \(n\) interlocutor. There are two general approaches to how higher order reasoning is set up. The first assumes that agents are strictly rational utility maximizers (e.g., Benz and van Rooij 2007, Jäger 2007b, Franke 2009). The second approach allows for some slack in agents’ tendency toward utility maximization (e.g., Goodman and Stuhlmüller 2013, Franke and Jäger 2014). As motivated below, we will follow the latter approach. For a finite state space this behavior is defined as follows for player \(i\):

\[
\rho_{n+1}(s|m; pr^i; L) \propto \exp(\lambda \frac{\sigma_n(m|s; L) pr^i(s)}{\sum_{s'} \sigma_n(m|s'; L)pr^i(s')});
\]

\[
\sigma_{n+1}(m|s; L) \propto \exp(\lambda(\rho_n(s|m; pr^i; L) - c_\sigma(m)))).
\]

That is, instead of using only the literal meaning of messages, higher order reasoners use Bayes’ rule to weigh their possible actions based on a conjecture about their interlocutor’s behavior. In both cases choice is regulated by a soft-maximization parameter \(\lambda \geq 0\) (Luce 1959, Sutton and Barto 1998). As \(\lambda\) increases choices approach strict maximization of expected utilities. For a sender this means that messages reasoned to have a high probability of being understood that are of low cost are increasingly prioritized over low success and/or high cost ones. In the case of receivers, states consistent with a conjecture of rational speaker behavior that are favored by their prior over states are more likely to be inferred. This so-called rationality parameter thereby allows for the representation of a range of behavioral strategies: from irrational behavior (\(\lambda = 0\)) up to the approximation of strictly rational behavior as \(\lambda\) approaches infinity.

In contrast to strict utility maximization, soft-maximization has the advantage of making explicit the degree to which rational behavior is necessary to explain pragmatic inference. Additionally, it allows us to consider cases in which some
deviation from optimal rational behavior might even be explanatory. From a more technical perspective, soft-maximization also has the desirable consequence that no choice is ever completely ruled out – even if acting upon it may be highly unlikely. By contrast, models that assume strict utility maximization require an additional assumption to deal with situations in which receivers reason that a message will never be sent (see Franke 2009, Franke and Jäger 2014, and Bergen et al. 2016:2ff for details and discussion). Such cases require attention because receiver strategies need to specify the behavior that ensues after the reception of any message. Otherwise, if, for one reason or another, such a surprise message is used, the receiver’s reaction to it would not be defined.

Another desirable consequence of not having degenerate choice probabilities will become relevant once iterated Bayesian learning comes into the picture (see Chapter 4 for details). That is, once we consider cases in which agents need not just use but first acquire language. In such situations non-zero choice probabilities allow naïve learners to entertain that all learning hypotheses could be compatible with the input they receive. Degenerate counterparts could, by contrast, lead to situations in which a language is faithfully transmitted solely because it harbors a single word/feature/construction that no user of a different language would ever use in virtue of strict utility maximization. Put differently, (at least minimal) variation in linguistic behavior opens up to the possibility of variation in the transmission of linguistic knowledge across generations.

Returning to single interactions and linguistic choice, as specified by (2.15) and (2.16), higher order reasoning types are assumed to behave rationally according to the expected behavior of a level $n$ interlocutor. This is a debatable design choice. For instance, a more flexible – and possibly more realistic – alternative would be for players to have beliefs about their interlocutor’s level of sophistication and for choice probabilities to be derived from these beliefs (see, e.g., Camerer et al. 2004). The assumption that players believe their interlocutor to be exactly one level less sophisticated than themselves is first and foremost made for simplicity. It primarily hinges on a trade-off between a more rigid reasoning procedure and increased model complexity. However, this assumption of myopic reasoning types has also been shown to succeed in predicting various empirically attested linguistic patterns (see Goodman and Frank 2016 for a recent overview). We will therefore opt for this simpler, and to my mind more perspicuous assumption, while bounding agents’ reasoning to a low degree of sophistication: level 1. This minimal degree of mutual reasoning will suffice to capture the classes of inferences we aim to characterize in this investigation.

To illustrate how higher order reasoning affects linguistic behavior, let us turn to the some-all context model again. With $\lambda = 1$ the choice probabilities of level-1
players are as follows:

\[
\sigma_1(\cdot | \cdot) \approx \begin{bmatrix}
m_{\text{some}} & m_{\text{all}} \\
0.62 & 0.38 \\
0.38 & 0.62 \\
\end{bmatrix}
\]

\[
\rho_1(\cdot | \cdot) \approx \begin{bmatrix}
m_{\text{some}} & m_{\text{all}} \\
0.58 & 0.42 \\
0.27 & 0.73 \\
\end{bmatrix}
\]

By contrast, with \( \lambda = 10 \) we have:

\[
\sigma_1(\cdot | \cdot) \approx \begin{bmatrix}
m_{\text{some}} & m_{\text{all}} \\
0.99 & 0.01 \\
0.01 & 0.99 \\
\end{bmatrix}
\]

\[
\rho_1(\cdot | \cdot) \approx \begin{bmatrix}
m_{\text{some}} & m_{\text{all}} \\
0.97 & 0.03 \\
0 & 1 \\
\end{bmatrix}
\]

The intuition behind the stronger association of \( m_{\text{some}} \) with \( s_{\exists \neg \forall} \) and that of \( m_{\text{all}} \) with \( s_{\forall} \) follows the same rationale given earlier. It results from agents’ mutual reasoning about (boundedly) rational language use.

More generally, in the some-all context model the association of \( s_{\exists \neg \forall} \) with \( m_{\text{some}} \) and that of \( s_{\forall} \) with \( m_{\text{all}} \) approaches 1 with \( \lambda \geq 17 \) for level-1 reasoning. For level-2 reasoning this result already obtains for \( \lambda \geq 7 \). Overall, the pragmatic strengthening of an underspecified message, in this case \( m_{\text{some}} \) to signal \textit{some but not all} rather than \textit{some and maybe all}, is predicted provided (i) some degree of rationality in choice (high \( \lambda \)) and (ii) some degree of mutual reasoning (\( n \)-level reasoning with \( n \geq 1 \)). Importantly, higher level reasoning can only lead to this strengthening if fueled by a common belief in rational language use. As suggested above by \( \lambda = 1 \), in cases of low rationality choice probabilities instead approach .5 as reasoning levels increase.

**Variation across models**

Models of rational language use share many fundamental components: they represent mutual reasoning as a reasoning hierarchy and characterize pragmatic inference as based on explicit representations of beliefs, preferences, and semantic conventions. The main difference across approaches lies in the choice functions of reasoning types and the depth of the reasoning on which they build. Four influential models that differ in this respect are the Rational Speech Act (RSA) model (Frank and Goodman 2012, Goodman and Stuhlmüller 2013), the Optimal Assertions (OA) model (Benz 2006, Benz and van Rooij 2007), the Iterated Best Response (IBR) model (Jäger 2007a, Franke 2009, Franke and Jäger 2014), and the Iterated Quantal Response (IQR) model (Franke and Jäger 2014).

In contrast to OA as well as to IBR and IQR, the RSA model was originally formulated as a hearer-centric model. Its main focus accordingly laid in capturing pragmatic interpretation rather than speaker behavior. The RSA model defines level-0 behavior and level-1 sender behavior in a similar fashion to definitions (2.13) and (2.16). However, in contrast to the above, no literal sender
is defined. Instead, pragmatic interpretation is identified with a level-2 receiver who reasons about level-1 sender behavior. Put differently, early instances of the model focused exclusively on the reasoning chain that defines level-2 receiver behavior. There are two other major differences to our definitions. First, only level-1 senders are assumed to soft-maximize. A receiver’s pragmatic inference instead only involves the inversion of level-1 speaker behavior using Bayes’ rule. Second, utilities in RSA models are associated with an information-theoretic measure of a message’s informativity about the sender state. For player $i$ this gives the following RSA-style behaviors:\(^3\)

\begin{align}
\rho_0(s|m; pr^i; L) &\propto L_{[s,m]} pr^i(s); \\
\sigma_1(m|s; L) &\propto \exp(\lambda(\log \rho_0(s|m; pr^i; L) - c_\sigma(m))); \\
\rho_2(s|m; pr^i; L) &\propto \sigma_1(m|s; L) pr^i(s).
\end{align}

In a nutshell, RSA views the sender’s goal as that of inducing a belief about the state in the receiver (with minimal effort if message cost is involved) by sending informative messages. The receiver’s goal is to form this belief.

Beyond numerical differences that can stem from the different ways in which level-1 sender behavior is defined, there are also conceptual differences between the RSA-approach and game-theoretical ones, embodied here by OA, RSA and IQR models. In the context of this investigation these differences are slightly obscured by the fact that we focus on interpretation games. To illustrate the two conceptions more clearly, recall that, in a general form, receiver strategies are defined as mappings from messages to (distributions over) acts. As mentioned earlier, acts can be interpretations; but they could equally well be reactions to animal alarm calls, such as hiding in a tree or hiding in a bush, or any other physical action such as passing the salt or bringing the keys. Under a game-theoretical view, utilities are preferences over state-message-act triples.

Following Qing and Franke (2015) we can call the RSA-view on utility belief-oriented and the game-theoretic view action-oriented. Under the former view, speakers care to maximize $\log \rho_0(s \mid m) - c_\sigma(m)$. That is, first and foremost, they care about the receiver’s beliefs and the receiver cares to form this belief. Under the latter view, receivers care to maximize their (expected) utility. It is their preference over state-(message-)act triples that matters.

The signaling behavior defined by OA, IBR, and IQR is more similar to definitions (2.13) – (2.16). Coming from a game-theoretic tradition, all three models are action-oriented as well. As for their differences, instead of soft-maximization, OA and IBR assume strict utility maximization at each step of the reasoning.

\(^3\)As with other models of rational language use, there is substantial variation within the RSA tradition. For instance, on whether receivers take senders’ preferences over messages into consideration; or on whether receivers’ priors only come into play at reasoning level 2, letting the prior of level-0 receivers instead be uniform across contexts. These finer differences do not need to concern us here, but see Qing and Franke (2015) for detailed comparison.
Chapter 2. Signaling Games: Analysis and Interpretation

hierarchy. IQR assumes soft-maximization at every step. This contrasts with our definitions of literal behavior in (2.13) and (2.14), where no tendency toward utility maximization is assumed. IBR and OA differ in two respects (Franke 2008). First, OA limits the receiver's reasoning sequence to $\rho_0-\sigma_1-\rho_2$ in a similar fashion to RSA. Second, prior state probabilities do not come into play in the OA model. Of course, behaviorally this difference disappears if the prior is uniform.

These variations can be seen as reflecting the explanatory goal of each model. OA, IBR, and IQR come from a game-theoretic tradition and aim to capture the signaling behavior of both players to understand their joint outcome. OA and IBR follow the classic assumption that choice is strictly rational. IQR accommodates a more diverse range of behaviors in a trade-off between some analytic results of OA and IBR and the possibility to capture finer-grained predictions involving different degrees of rationality. This ability makes IQR models more attractive for investigations that involve empirical data (Degen and Franke 2012, Franke and Jäger 2016a; see Chapter 3 for a concrete application). In its original formulation, the RSA model is more constrained in that it focuses only on a particular kind of receiver behavior. This constraint is a consequence of its original goal to model experimental data involving only pragmatic interpretation. However, nothing prevents the model from being extended to characterize sender behavior as well (see, e.g., Franke and Degen 2016 for such an extension). In sum, I believe it is fair to say that the commonalities among these approaches far supersed their differences. Within any approach it is common to find a number of slight differences in how each analysis sets up the linguistic behavior of agents. However, these changes mainly hinge on the phenomenon to be explained. They do not purport to be fundamentally new proposals in their own right.

As for my own assumptions on these matters, the definitions in (2.13)–(2.16) and their instantiations in the following chapters are mainly motivated by the advantages conferred by soft-maximization in contrast to strict utility maximization mentioned above. As for other differences, contra IQR, level-0 soft-maximization is not assumed in cases where literal behavior is not considered to correspond to actual linguistic behavior. Contra RSA, higher level receivers are assumed to soft-maximize because their goal is not only to form a posterior belief about the sender state, but rather to maximize the utility derived from the act they perform upon reception of a message. That is, I will follow the game-theoretic view of action-oriented signaling. On this matter we should note that whether there is a conceptual difference between belief- and action-oriented models depends on the situation at hand. In cases in which interlocutors care about belief-formation they evidently coincide (see van Rooy 2004b for detailed comparison of notions of utility and related information-theoretic measures). In terms of technical differ-

\[\text{In Chapter 4 we will come to contrast literal behavior with behavior that results from higher order reasoning under evolutionary dynamics. In this setup literal behavior is taken to correspond to behavior that can actually be adopted. Consequently we assume it to come with a rationality parameter that regulates choice (see Franke and Degen 2016 for a similar choice).}\]
2.5. Final Remarks

This chapter touched upon a number of different questions. The two central ones are:

(i) How can linguistic behavior be represented?

(ii) How can linguistic outcomes be analyzed?

On question (i) we follow the view of language use as a strategic endeavor of information transfer and employ signaling games as context models of features relevant to an interaction: individuals’ preferences and beliefs, together with the linguistic material relevant to the situation at hand. At a general level, this view of communication identifies linguistic knowledge with operational knowledge (Parikh 1994:530ff). What is relevant for communication is not that speaker and receiver share a language. Rather, what is relevant is that their language use effects successful behavior. In the case of interpretation games this translates to an agreement between a sender’s information state and a receiver’s interpretation. Importantly, this view allows for mismatches in interlocutors’ subjective priors (Chapter 3) or even their semantics (Chapter 4 and 5).

As for question (ii), we argued that, where possible, linguistic outcomes should be explained in terms of the processes that lead to their emergence and stability rather than by appeal to their optimality. Particular phenomena will of course call for the analysis of particular processes, at possibly distinct levels of interaction. Irrespective of these differences, the common thread of this investigation lies in that they build on models of rational language use at the individual level. This allow us to inspect the joint role of semantic conventions and language use in the emergence of pervasive phenomena at the semantics-pragmatics interface.