Cognitive Science 34 (2010) 583–601

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2010.01095.x

The Evolution of Relevance

Thomas C. Scott-Phillips

School of Psychology, Philosophy, and Language Sciences, University of Edinburgh

Received 5 June 2009; received in revised form 8 December 2009; accepted 10 December 2009

Abstract

With human language, the same utterance can have different meanings in different contexts. Nevertheless, listeners almost invariably converge upon the correct intended meaning. The classic Gricean explanation of how this is achieved posits the existence of four maxims of conversation, which speakers are assumed to follow. Armed with this knowledge, listeners are able to interpret utterances in a contextually sensible way. This account enjoys wide acceptance, but it has not gone unchallenged. Specifically, Relevance Theory offers an explicitly cognitive account of utterance interpretation that presents a radical challenge to the neo-Gricean paradigm. Evolutionary considerations are one way in which we can choose between competing theories. A simple game-theoretic model of the evolution of communication is presented, and it is used to derive a number of basic qualities that will be satisfied by all evolved communication systems. These qualities are observed to precisely predict the foundational principles of Relevance Theory. The model thus provides biological support for that enterprise in general, and for the plausibility of the cognitive mechanisms that it describes in particular.

Keywords: Linguistics; Psychology; Biology; Communication; Pragmatics; Evolution; Relevance

1. Introduction

Ambiguity is commonplace and indeed inevitable in everyday language. An utterance produced in one context may have a quite different meaning in a different context (Atlas, 2005; Austin, 1955; Carston, 2002; Grice, 1975; Sperber & Wilson, 1995). Despite this, listeners almost invariably converge upon the correct meaning that the speaker intends to convey. How is this achieved? The still widely accepted answer to this question was provided by Grice (1975), who posited the existence of a *cooperative principle*, which comprises four *maxims of conversation*: Quality (tell the truth), Quantity (do not say too

Correspondence should be sent to Thomas C. Scott-Phillips, School of Psychology, Philosophy and Language Sciences, University of Edinburgh, Edinburgh, UK EH8 9AD. E-mail: thom@ling.ed.ac.uk

much or too little), Relation (be relevant), and Manner (be clear and concise). It is, according to Grice, because listeners assume that speakers follow these maxims that they are able to interpret the literal meaning of an utterance in a contextually sensible way.

Since Grice's seminal contributions, numerous refinements, additions, and extensions of his work have been proposed (e.g., Horn, 1984; Leech, 1983; Levinson, 1983, 2000). However, the Gricean foundations remain widely accepted (Levinson, 1989), and the cooperative principle is presented as the established paradigm in all introductory textbooks on pragmatics. This acceptance means that the neo-Gricean framework has also been influential in associated disciplines such as psycholinguistics (e.g., Clark, 1996) and the philosophy of language (e.g., Lycan, 2008). One alternative is *Relevance Theory* (RT; Sperber & Wilson, 1995), which takes an explicitly cognitive approach and supplants the four maxims with a single notion of *relevance*. Its success in this regard has been mixed; although RT has its adherents (see Yus Ramos, 1998), neo-Gricean frameworks remain dominant.

How should one choose between these competing explanatory frameworks? Theoretical argumentation from either side is typically grounded in philosophy, theoretical linguistics, or, in the case of RT, cognitive science. There is a growing interest in experimental approaches to pragmatics (Noveck & Sperber, 2004), but it is too early for this work to offer firm conclusions on the major theoretical questions. This study introduces a new factor that can be used to think about utterance interpretation: biological evolution. The capacity for successful utterance interpretation is a psychological and cognitive phenomenon, and as such is part of our biological make-up. Our accounts of how listeners converge upon correct speaker meanings should therefore posit psychological mechanisms that are, at a minimum, consistent with evolutionary theory. Better still, evolutionary theory may be able to describe some basic properties that such an evolved system might possess. In short, compatibility with evolutionary theory is one criterion by which we can differentiate between theoretical frameworks within linguistics (Kinsella, 2009). This study uses that constraint to argue in favor of RT.

The next section introduces the issues at hand in more detail. It explains why utterance interpretation is not a trivial problem and outlines both the Gricean and RT approaches. A very simple game-theoretic model of the evolution of communication, which describes some basic qualities that evolutionarily stable communication systems will necessarily satisfy, is then developed. Those qualities are then found to precisely map onto the founding principles of RT, and as such they provide biological support for that enterprise. Possible objections to the analysis presented here are also discussed.

2. Pragmatics

2.1. Ostension, inference, and the cooperative principle

Broadly speaking, we can identify two approaches to the study of linguistic communication. Within the *code model*, communication is conceived of as a process whereby two information-processing devices (human brains) directly map internal meanings into external

signals in both production and reception. In this picture, to encode or to decode an utterance is to perform an act of machine translation, in which a lexicon is searched for the meaning of each of the utterance's constituents, and these meanings are then combined to form the meaning of the utterance. A similar process, in reverse, accounts for production. The defining formulation of this model is Shannon and Weaver's (1949) *Mathematical Theory of Communication*, which sowed the seeds of information theory and is still the dominant paradigm of communication in artificial intelligence and associated disciplines. More specifically, there is a wide if implicit assumption among many linguists and cognitive scientists that this is a reasonable way in which to conceptualize communication.

Over the past 40 years or so pragmatics has developed and refined an alternative to this picture (e.g., Austin, 1955; Grice, 1971, 1975; Schiffer, 1972; Sperber & Wilson, 1995). The *ostensive–inferential* model of communication posits that communication is achieved through the production and interpretation of evidence for the meaning that is to be communicated. The act of production is called ostension and the act of comprehension inference. The evidence is provided through the physical alteration of the shared environment (i.e., by speech, gestures, or whatever other medium is used), an act that triggers the inference of the intended meaning. For example, if I offer my girlfriend a cup of tea in the morning, she may respond, ''I've already cleaned my teeth.'' In doing so, she provides evidence that she wishes to decline my offer, yet she does not say as much explicitly. Neither must the evidence necessarily be verbal, nor even linguistic: I can gesture toward a friend's newly arrived plate of chips, and in doing so provide evidence of my desire to have one of the chips myself. Different pragmatic theories (see Huang, 2007; Levinson, 1983, for surveys) disagree about precisely how this communication is achieved, but all agree that production is ostensive and that its goal is to induce a particular change in the listener's mind, and that comprehension is inferential and its goal is to discover the speaker's intended meaning.

These facts bring with them the problem of linguistic underdeterminacy: the (often huge) gap between the literal meaning of an utterance (called linguistic meaning) and the meaning that is actually intended (called speaker meaning) (Atlas, 2005; Austin, 1955; Carston, 2002; Grice, 1975; Sperber & Wilson, 1995). How do listeners plug this gap? The classic solution (Grice, 1975) proposes that this underdeterminacy is resolved by the *cooperative principle* described in Section 1. There may be several different ways for speakers to provide the evidence necessary to lead the listener to the correct intended meaning. Grice suggested that speakers will in general choose the way that is most consistent with these maxims. Because of this, listeners can assume that speakers have followed the cooperative principle, and they use this knowledge to converge upon some contextual sensible information. Consider, for example, the following exchange (Levinson, 1983):

A: Where's Bill?
B: There's a yellow VW outside Sue's house.

Taken literally, B's utterance fails to address A's question. It appears to violate the maxims of Quantity (the presence of the yellow VW is more information than was requested) and Relation (what has the yellow VW to do with where Bill is?). There is, then, an apparent

failure of cooperation. However, rather than draw this conclusion, the listener assumes that this is not what has occurred and searches instead for some nonliteral interpretation of B's utterance which does satisfy the four maxims.

Most subsequent developments in pragmatics (e.g., Horn, 1984; Levinson, 1983) have altered, refined, and fine-tuned this basic picture. Consequently, neo-Gricean pragmatics resembles ''an untidy collection of usage principles, accrued over decades of careful observation, which together give some substantial account of uncoded utterance meaning. It may be a bit ramshackle, but it delivers the goods; and new developments help to remedy deficiencies'' (Levinson, 1989, p. 469). It is in this context that RT can be seen as ''an ambitious bid for a paradigm-change in pragmatics'' (Levinson, 1989, p. 469).

2.2. Relevance Theory

This section can, of course, only offer a very brief sketch of the central ideas of RT, but it is intended to contain sufficient detail for the purposes of this study. The interested reader is referred to the original text (in particular the second edition, which is the definitive statement of the theory; Sperber & Wilson, 1995) or one of the numerous précis (e.g., Sperber & Wilson, 1987; Wilson & Sperber, 2004).

Motivated by dissatisfaction with the cognitive aspects of the (neo-)Gricean approach to communication (Sperber & Wilson, 1995), RT supplants the four Gricean maxims with a single notion of *relevance*. This is defined as a tradeoff between two competing properties of an utterance. On the one hand, there are the worthwhile changes in the receiver's representation of the world that are warranted by the utterance. These are called *positive cognitive effects* and include the strengthening, weakening, or elimination of previously held knowledge and the provision of premises from which to infer new knowledge. For example, if a colleague's utterance leads me to a better understanding of some problem, by any of these means, then it has had a positive cognitive effect upon me. As such, positive cognitive effects can be seen as the payoff associated with correct utterance interpretation. To achieve these cognitive effects, some time and energy must be expended. This *processing effort* must be weighed against the cognitive effects of an utterance, on the other hand. The result gives us a measure of relevance; see Fig. 1.

Thus, all other things being equal, the greater the positive cognitive effects achieved by processing an utterance, and the lower the processing effort expended in processing the utterance, the greater the relevance of that utterance. This is how relevance is defined within RT. This means, of course, that the same utterance will have different degrees of relevance to different individuals and at different times: The listener's prior knowledge will impact on both the degree of positive contextual effect achieved by an utterance and the amount of processing effort required to comprehend it.

Relevance Theory then claims that we make sense of ambiguous utterances by assigning to them the interpretation that maximizes their relevance. More precisely, RT claims that the very production of an utterance raises in listeners an expectation that the utterance is of relevance, and that this expectation, coupled with the evidence provided by the linguistic meaning of the utterance, is sufficient to guide them to the correct speaker meaning. This

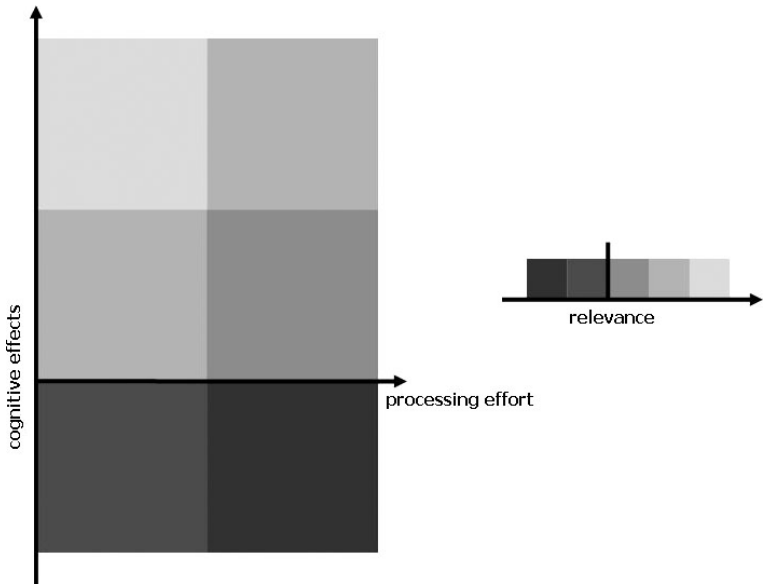

Fig. 1. Relevance of an utterance to an individual. Within Relevance Theory, utterances are said to be relevant if they achieve positive cognitive effects (i.e., they cause a worthwhile change to an individual's representation of the world). Weighed against this is the processing effort required to achieve such effects—if two utterances achieve the same effects but one with less processing effort than the other, then they will have different degrees of relevance. Here, these two measures are plotted against one another, and the degree of relevance is indicated by shades of gray: the lighter the shade, the more relevant it is. Irrelevant utterances are colored in particularly dark shades.

vision is captured by two empirical claims about how human cognition processes utterances. These *principles of relevance* form the theoretical foundation of RT, upon which the rest of the theory relies.

The first (the *cognitive principle of relevance*) is that *human cognition tends to be geared toward the maximization of relevance*. This means that, when listeners process a stimulus, they will converge upon the interpretation that grants the stimulus the maximum degree of relevance. The second (the *communicative principle of relevance*) is that *every utterance*[1] *carries a presumption of its own optimal relevance*. This means that the very production of an utterance raises in listeners an expectation that the utterance is relevant enough to make it worth the listener's while to process it; and, moreover, that the utterance is the most relevant one compatible with the speaker's abilities and preferences.

Contrary to a common assumption (e.g., Clark, 1996; Kopytko, 1995), RT does not reduce the four maxims to one. The theoretical status of the principles of relevance within RT is quite different to the theoretical status of the four maxims. Whereas the latter are behaviors that speakers are simply thought to aim for, the former are argued, on the basis of a number of simple observations, to be fundamental facets of human communication and cognition. Unlike the cooperative principle, the principles of relevance are inviolable. The four maxims of conversation ''are only operable on the back of considerable amounts of

prior context-dependent inference'' (Wedgwood, 2005, p. 49). The principles of relevance, in contrast, attempt to explain how that context-dependent inference can occur in the first place. This is why RT can be seen as a radical challenge to, rather than a refinement of, the Gricean paradigm.

If this is correct, then the Gricean observations about how humans behave in conversation, which are captured by the maxims of conversation, should be describable in terms of the principles of relevance. It is not difficult to see how this could be done. The maxim of Quantity is satisfied as saying too little will not maximize cognitive effects, and saying too much demands increased processing effort for no additional cognitive effects. The maxim of Manner is satisfied as clarity and ease of comprehension speak to the minimization of processing effort. The maxim of Relation is satisfied practically by definition. The maxim of Quality is slightly more complex case, but it is still subsumed by the principles of relevance (Wilson & Sperber, 1981, 2002): Speakers who wish to maximize cognitive effects could say that for which they do not have sufficient evidence, but if they do, then they will lose the opportunity to induce more cognitive effects at a later date, since the listener will lose trust in them. The various neo-Gricean approaches (e.g., Horn, 1984; Levinson, 1983) can be accounted for in a similar way. As such, then, RT should not be seen as a reduction of the four maxims to one. Rather, the principles of relevance are argued to be fundamental features of human communication that in turn give rise to (neo-)Gricean observations about conversation.

The goal of this study is to show that the functional role performed by the two principles of relevance is an inevitable property of all evolved communication systems. To do this, the next section describes a very simple mathematical model of the evolution of communication, which is used to derive some basic statements about the functionality of signals and their interpretations.

3. Evolution and communication

3.1. Different types of cooperation in communication

The model of evolution of communication presented below is not a classic signaling game in which one player (the signaler) has some knowledge about the state of the world and must choose what signal to send to the other player (the receiver), who does not have that knowledge. Under such circumstances, the receiver cannot know the exact payoff that she will receive for any given signal–response pair, as that payoff also depends upon the state of the world, which is unknown to her. Games of this type are said to be games of *incomplete information*. The model presented below, in contrast, uses a game of *complete information*—both participants know the payoffs that will result from each combination of possible behaviors. Initially, this seems inappropriate: If there is no state of the world about which the signaler is trying to communicate, then why is there any communication at all? However, this model attempts to address a different problem from that of previous models.

To see why, we must distinguish between the different types of cooperation involved in communication (Scott-Phillips, 2010). In two different senses, communication is an inherently cooperative act. First, the meanings of signals must be consistent across individuals. If signal *A* carries meaning *X* for one organism but meaning *Y* for a different organism, then when the first organism produces signal *A* to communicate *X*, it will fail since the receiving organism will interpret the signal as *Y*. Without this foundation successful communication cannot occur. We can term this *communicative cooperation.* Second, those signals must, on average, be honest: If they are not then receivers will cease to have trust in them, and the system will collapse. We can term this *informative cooperation*. In a third sense, however, the ends to which communication is employed may be cooperative or competitive: We may work toward mutually beneficial goals, or there may be a conflict of interests. We may term this *material (non)cooperation*. These different types of cooperation are summarized in Table 1.

Evolutionarily stable communication need not exhibit material cooperation. Communicative and informative cooperation, however, are necessary conditions for stability. The second of these, informative cooperation, is concerned with the honesty of signals, and it is the defining problem of animal signaling theory (Maynard Smith & Harper, 2003; Searcy & Nowicki, 2007). As such it has been a central focus of previous models of the evolution of communication. The signaling games described above are the appropriate game-theoretic tool with which to approach it. However, our goal in this study is not to investigate the honesty of signals, but rather the question of what interpretation a listener should grant to an utterance when faced with ambiguity. The corresponding problem is thus not the matter of informative cooperation but rather the matter of communicative cooperation: How do signalers and receivers agree on the meaning of a particular signal? Crucially, the two players' interests are aligned here. If they fail to converge upon a common understanding of a signal, then they cannot communicate with one another for any ends, informatively cooperative or not. The appropriate game, then, is not one *of* communication, in which the signaler attempts to communicate to the receiver some previous unknown fact about the world; but rather one *about* communication, in which signaler and receiver attempt to align with one another.

Table 1
The different types of cooperation involved in communication

Type of Cooperation	Gloss	Necessary for Evolutionarily Stable Communication?
Communicative	Do interlocutors have the same meaning-form mappings as each other?	Yes
Informative	Does the signal carry reliable information—is it honest?	Yes
Material	Is communication being used to achieve mutually beneficial goals?	No

3.2. A general model of the evolution of communication

The model, which is entirely general (i.e., not specific to humans), will produce three basic statements about what it means to achieve communicative cooperation. Two of these are evolutionary game-theoretic axioms: that the two participants, signaler and receiver, will seek to maximize their payoffs. The third is that at equilibrium signals will not just be beneficial to the signaler but also to the receiver. The argument is really an extremely simple case of backwards induction, and the main point can be grasped quite straightforwardly: If signals are not beneficial to receivers, then they will evolve not to attend to the signals at all, and the system will consequently collapse. As a result, at equilibrium, the very production of a signal reveals that it is worth the listener's while to process the signal. Despite its intuitiveness, it is useful to express the argument more formally. Mathematical proof is provided in the Appendix.

We begin with the simplest version of the model, with just two possible actions, *A* and *B*, and two possible reactions, *X* and *Y*, each combination of which constitutes a different interaction. In essence, actions are utterances and reactions are interpretations. The subsequent interactions will result in payoffs for the actor and reactor, and can thus be mapped onto a graph, with the payoff to the actor on the *x*-axis and the payoff to the reactor on the *y*-axis, as per Fig. 2. The payoff to the actor will be termed the act's *impact*, and the payoff to the reactor will be termed its *pertinence*. As such, impact is a measure of how worthwhile it is for the signaller to produce the signal, whereas pertinence is a measure of how much the receiver has to gain from a signal. It is quite conceivable that one could be high and the other low. For example, suppose that I have a secret that I do not wish to reveal to you, but in which you would be very interested. You then stand to gain significantly from the utterance (i.e., high pertinence), but it is not at all worthwhile for me to reveal the secret (i.e., negative impact).

The question to be asked is: Which of these interactions is evolutionarily stable? An initial reaction is that interactions closest to the top-right corner will result, as that maximizes the payoffs to both participants, but the matter is not quite so simple. An important fact about communication is that receivers react to signalers; the two behaviors are sequential. This means that this game is different from the prisoner's dilemma, the stag hunt, and several other games, in which the moves are simultaneous. The present game is thus *dynamic* rather than *static*. The actor moves first and the reactor, *knowing what choice the actor has made*, moves second. This means that the reactor is guaranteed to achieve the optimal payoff given the actor's behavior, and also that the actor's payoff is contingent upon this. This is not the case in static games, and it is a fact that the actor must take into account if she is to maximize her payoffs. In short, she should calculate what the reaction would be for each possible action and then choose her action accordingly. There is, then, a feedback loop, in the sense that the expected behavior of the reactor is an input into the behavior of the actor; and once that behavior is performed then the reactor simply chooses the reaction that is most pertinent (recall that pertinence is defined as the payoff to the reactor). The action–reaction pair that results is known as the *Stackelberg equilibrium*.[2] The Stackelberg equilibrium is not the Bayesian equilibrium of classic signaling games. As discussed above, this is not a game in which the signaler has some knowledge about the state of the world and must choose what

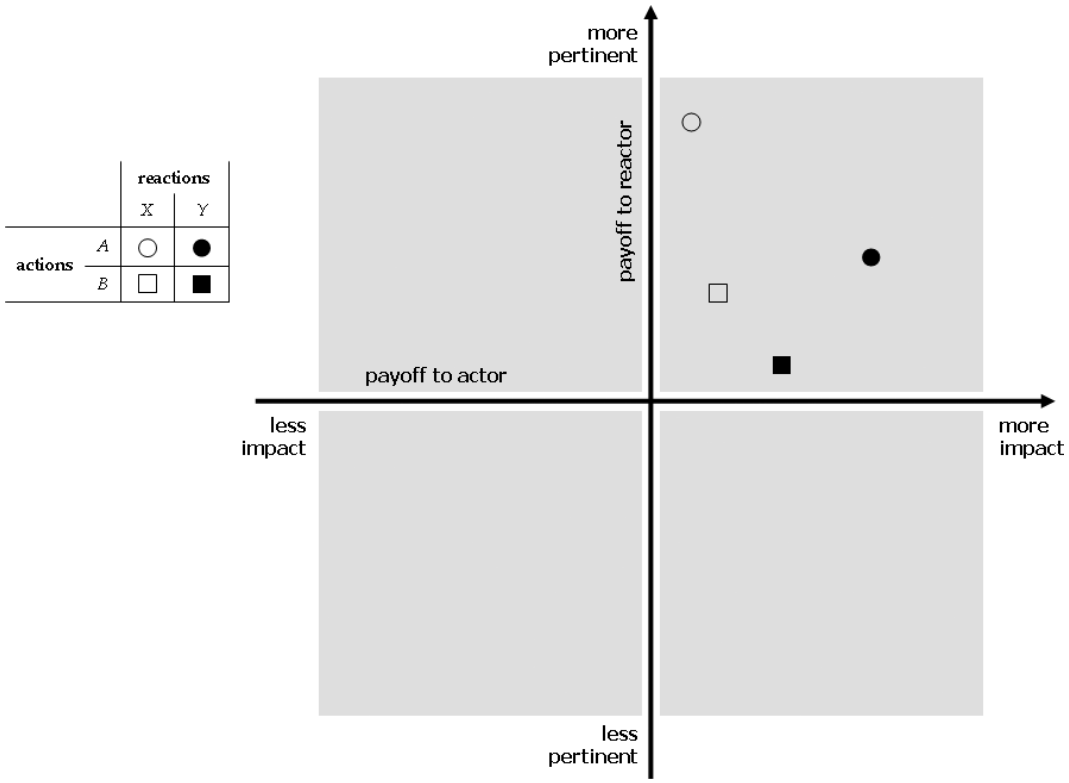

Fig. 2. Four possible communicative interactions. The actor can perform either action *A*, which will result in one of the two circles, or action *B*, which will result in one of the two squares. The reactor can perform either reaction *X*, which will result in one of the clear shapes, or reaction *Y*, which will result in one of the filled shapes. The filled circle is closest to the top-right corner of the graph, and thus maximizes the net payoffs achieved by the participants, but we will see that it is not evolutionarily stable.

signal to send to the receiver. Instead, the receiver can, in effect, choose what payoff he will receive given the signaler's behavior, as he can decide what interpretation is most warranted by the signal. As discussed, this is appropriate for a game-theoretic analysis of communicative (rather than informative) cooperation.

Returning to Fig. 2, we can observe that the actor should take note of which instance of each shape has the most impact (recall that impact is defined as the payoff to the actor) and then choose from this reduced set of possible actions. This is depicted in Fig. 3, where the two possible outcomes for each action are grouped together. It should be apparent that if the actor performs action *A* (and so chooses the circles) then the reactor will opt for the clear circle, as that offers a greater payoff; and if the actor performs *B* then the reactor will choose the clear square. The actor thus has a choice not of all four outcomes but of the two clear shapes, and of these it is the square that offers the greater payoff. This is thus the Stackelberg equilibrium, even though the filled circle would represent a greater payoff to both participants (i.e., would be Pareto optimal); see Fig. 4.

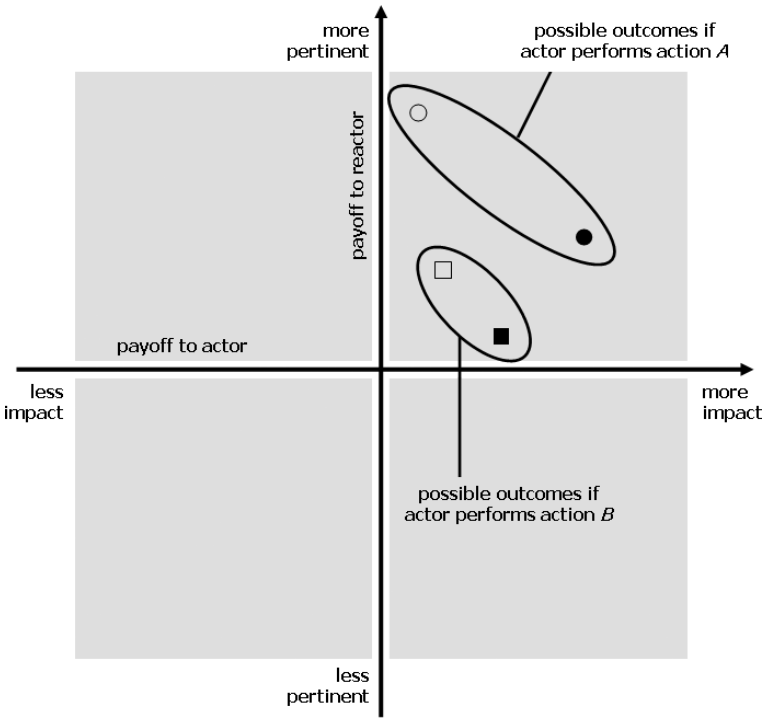

Fig. 3. The actor's choice. Following Fig. 1, either the actor performs *A*, which will result in one of the interactions marked by a circle, or performs *B*, which will result in one of the interactions marked by a square.

Missing from this model is what would happen if the Stackelberg equilibrium were not in the top-right quadrant of this graph; that is, if the interaction results in a negative payoff for either participant. Under such circumstances, we should expect the participants who incur the negative payoff not to partake in the interaction at all. To capture this, we should add to the model the possibility of doing nothing. To do this, we simply include an additional pair of behaviors, which we term the *null action* and *null reaction*. These appear in Fig. 5, alongside four other interactions that share the same relative status to each other as the interactions in previous figures. The triangle represents the outcome if the actor does nothing: Neither participant receives a payoff at all, either positive or negative. The hashed circle and square just to the left of the origin represent the outcome if the reactor does nothing (i.e., if they ignore the actor)—the negative payoff to the actor reflects the small but nonzero energy and opportunity costs of the act. We should, if there were no null behavior, expect the equivalent interaction to be the Stackelberg equilibrium. However, the possibility that the participants can opt out of the interaction altogether has changed the equation: It is plain to see that, in this new model, the Stackelberg equilibrium is the null interaction represented by the clear triangle. As such, the system has now collapsed, as there are no interactions that are worth the participants' while: The equilibrium state is one in which there is no action, and thus no reaction either.

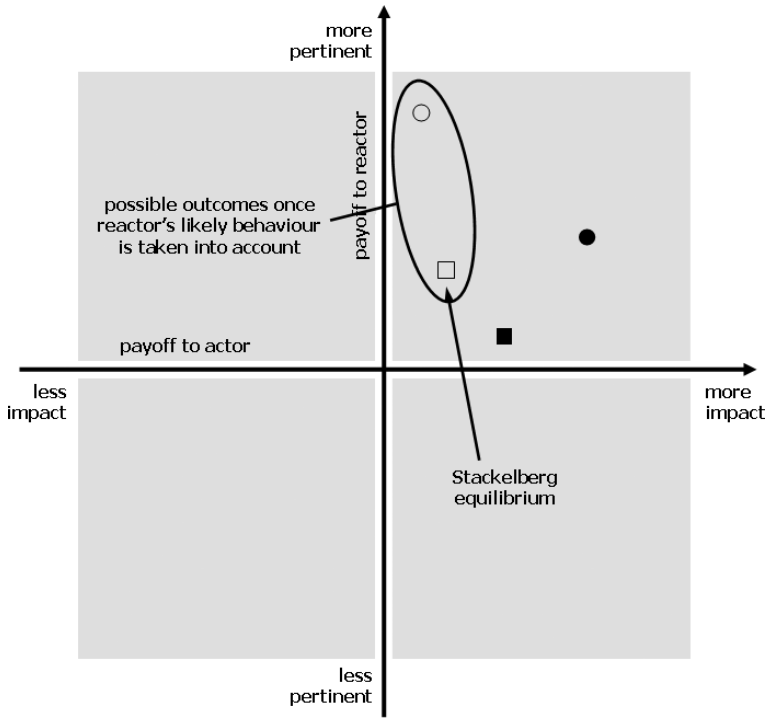

Fig. 4. The Stackelberg equilibrium. To maximize her own payoffs, the actor should account for her interlocutor's likely reaction. Here, the reactions that carry the greatest payoff to the reactor (i.e., are most pertinent) for each action are those depicted by clear shapes. The actor thus chooses between these clear shapes. The one with the greatest payoff to the actor (i.e., is has the most impact) is the Stackelberg equilibrium.

A comparison of Figs. 4 and 5 (i.e., the scenarios in which doing nothing is and is not the Stackelberg equilibrium) shows that if the Stackelberg equilibrium includes neither the null action nor the null reaction, then the interaction has a positive payoff for both actor and reactor, for example, in the top-right quadrant of the graph. (If the possible outcomes are grouped in the top-left quadrant of the graph, then the actor will choose the null action; and if the outcomes are grouped in the bottom-right, then the reactor will choose the null reaction.) Moreover, as participants maximize their payoffs, the reaction will be the most pertinent reaction possible. This is the main result of the model, and as shown in the Appendix, it is entirely generalizable; that is, it applies not just to the simple version described thus far, with just two possible actions and reactions, but to a model in which the sets of possible actions and reactions are both infinite in size.

3.3. General principles of communication

Attention is drawn to three basic facts about this model. The first two are axiomatic: that both parties will seek to maximize their payoffs. For signalers, this means that they will

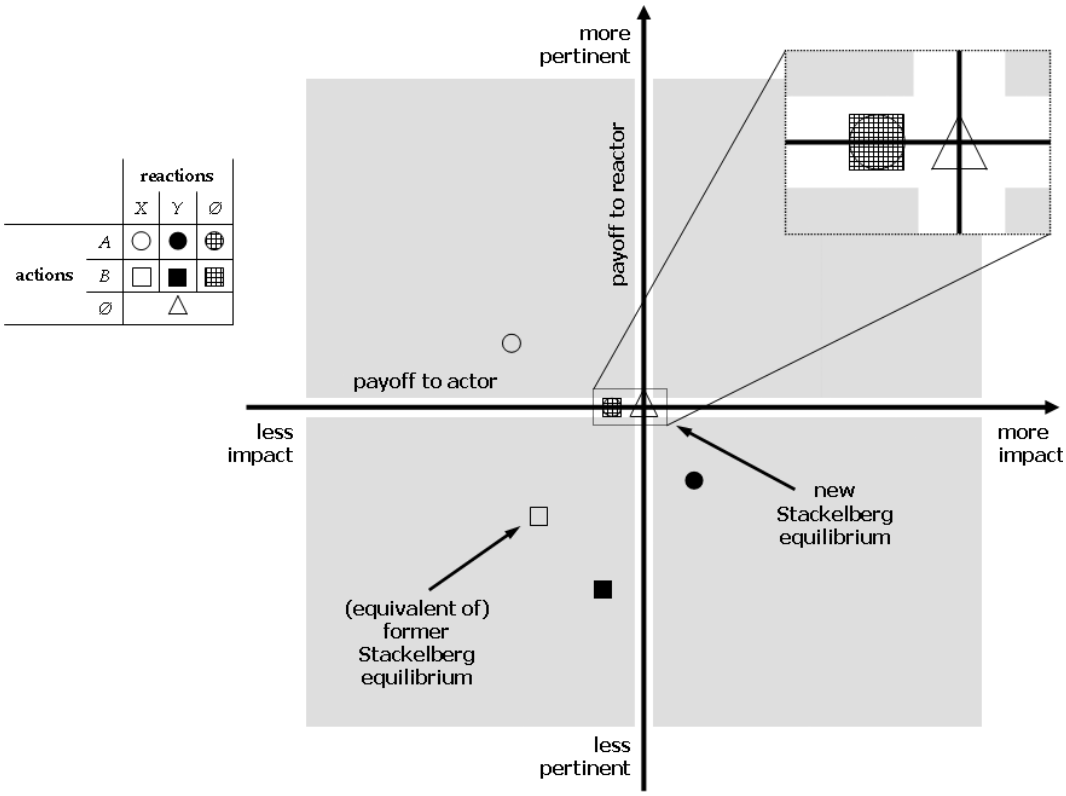

Fig. 5. The inclusion of null actions and reactions. The clear triangle represents the actor choosing to do nothing, and the hashed circle and square just to the left of the origin represent interactions where the reactor does nothing. The relative status of the other interactions has been maintained from Fig. 3, but moved to the bottom-left quadrant so that payoffs to both participants are negative. In this situation, the null action (the clear triangle) is now the Stackelberg equilibrium. (This is also true if the other interactions are in the top-left or bottom-right quadrants; see main text.)

choose the signal that is most likely to convey the meaning the signaler intends it to. As the payoff to signalers is expressed in terms of impact, we call this the *principle of maximal impact*—that *evolved signals will tend to be geared toward the maximization of impact*. Consider, as an example, bird song, which is used for both courtship and for territorial defense. All that this first axiom states is that a bird that wished to pursue a mating opportunity will produce courtship song, rather than territorial sing, as that is most likely to induce the intended reaction.

For receivers, payoff maximization simply means that they will choose the interpretation that maximizes the benefit they can receive from the signal. As the payoff to receivers is expressed in terms of pertinence, we call this the *first principle of pertinence*—that *listeners will grant to signals interpretations that maximize pertinence*. For the receiver of a songbird's courtship signal, one possible interpretation is that the bird is seeking to defend territory. However, given the principle of maximal impact, this conclusion is likely to be

wrong and will thus carry a negative payoff (as the receiver will now have an inaccurate representation of the world). In contrast, the conclusion that the songbird is seeking a mate may have a positive payoff.

These two facts are the model's axioms. The third fact to be highlighted is the main result of the model: that in an evolutionarily stable system then if a signal is produced, it will be maximally pertinent. We call this the *second principle of pertinence*—that *every signal carries a presumption of its own optimal pertinence*. This means that the very production of a signal is, in effect, a statement that the signal is worthy of the audience's attention, and moreover, that the intended interpretation is the one that maximizes the payoff to the receiver.

It should already be apparent that there are obvious similarities between these principles of pertinence and RT's principles of relevance. The next section discusses exactly how we should interpret this fact.

3.4. Relevance as a mechanism of linguistic communication

To allow for a proper comparison between the principles of pertinence and of relevance, all are repeated here:

First principle of pertinence
Listeners will grant to signals interpretations that maximize pertinence
First (cognitive) principle of relevance
Human cognition tends to be geared toward the maximization of relevance
Second principle of pertinence
Every signal carries a presumption of its own optimal pertinence
Second (communicative) principle of relevance
Every utterance carries a presumption of its own optimal relevance

That there are similarities here is immediately apparent. How should we interpret this fact? The two principles of pertinence are very basic statements about the evolutionary functionality of signals. Two are axiomatic of adaptationism; the third a consequence of the interdependence of signals and responses that is an inherent part of communication (Scott-Phillips, 2008). What can these statements tell us about the way in which humans achieve communication? Natural selection will tend toward mechanisms that deliver the closest fit to the evolutionary function (Fisher, 1930; Grafen, 2007; Hamilton, 1964). One particularly exact way in which to achieve that fit would be to have some proxy measure of pertinence that can be cognitively implemented. The suggestion, then, is that relevance performs this task in humans.

This claim should be fleshed out somewhat. Evolutionary biology makes a clear distinction between ultimate and proximate explanations (Mayr, 1963; Tinbergen, 1963; West, Griffin, & Gardner, 2007). The former are concerned with the evolutionary pressures that give rise to the trait or behavior in question, and as such explain *why* a particular trait exists; that is, they explain evolutionary functionality. The latter describe *how* those functional goals are achieved—the various psychological, physiological, physical, chemical, and other

phenomena that deliver such outcomes. These two types of explanation are not continuous with each other, nor should we choose between them. On the contrary, they are distinct from one another and complementary. To properly understand a behavior, we must account for both the evolutionary rationale for its existence and the various material phenomena that fully describe it. This framework has been widely adopted by evolutionarily minded psychologists (Barkow, Cosmides, & Tooby, 1992; Dunbar & Barrett, 2007), but its import has yet to be recognized within linguistics (Scott-Phillips, 2007).

The three principles of communication derived in the previous section (the two principles of pertinence above, and the principle of maximal impact) are functional explanations—they are concerned with the evolutionary logic that explains the existence of a certain class of behaviors. They will apply to all evolved communication systems. The evolutionary game-theoretic model presented here thus demonstrates that all organisms that communicate must have mechanisms by which these three principles of communication are achieved. These mechanisms may take many possible forms. In some cases, they may consist of simple, automatic causal processes. Some bacteria, for example, communicate by a process known as quorum sensing, in which a coordinated population response is controlled by diffusible molecules produced by individuals (see Diggle, Gardner, West, & Griffin, 2007, for a review). In other cases, the mechanisms will be more sophisticated. Specifically, human cognition is a particularly powerful and flexible tool. What the principles of communication tell us is that whatever the mechanism, it should be calibrated to enable speakers to perform online calculation of the potential impact that an utterance might have, and to enable listeners to maximize the cognitive effects of such utterances. If it is not calibrated in these ways, then it will fail to achieve the principles of communication, and hence will be selected against. The claim, then, is that the two principles of relevance describe the proximate mechanism by which the evolutionary functionality of communication is achieved in humans.

The model of the evolution of communication described in the previous section also produced a *principle of maximal impact*—that evolved signals will tend to be geared toward the maximization of impact. This means, first, that signalers will (on average) only produce signals when it is worth their while to do so; and second, that when they do they will produce signals that most effectively achieve the signal's function. Again, these points will be true of all evolved communication systems and are unsurprising—why should we expect a signaler to signal if there is no payoff to them of doing so? In humans, the reasons for doing so are many any varied, and they include social bonding (Dunbar, 1997), sexual activity (Burling, 2005; Miller, 2000), status enhancement (Dessalles, 1998), and others. If we assume that there is some cognitive cost associated with utterance production (in the same way that RT suggests that utterance interpretation requires some degree of processing effort), then the principle of maximal impact suggests that whatever the reason for the utterance, it will be produced in such a way so as to keep this cost as minimal as possible given the goal of the utterance. All in all, utterances will be designed to achieve the maximal cognitive effect at the minimal processing effort, as suggested by RT.

4. Possible objections

Four possible objections must be addressed before the conclusions of this work can be fully accepted. First, it might be claimed that although the principles of pertinence do map onto the principles of relevance quite directly, they are equally compatible with some version of the neo-Gricean paradigm; in other words, that the maxims of conversation (or indeed any version of the neo-Gricean paradigm) are equally able to achieve the functionality described by the principles of pertinence, and which is requisite for consistency with evolutionary theory. The argument may then proceed in one of two ways. The existence of the cognitive mechanisms described by RT (i.e., the tendency to maximize relevance, and the tendency to produce utterances that are optimally relevant for the listener) must either be accepted or denied. If they are accepted, then the Gricean maxims are straight forwardly explained as emergent behavioral properties of these mechanisms, in the way discussed in the section on pragmatics. The fact that they approximate the functionality of the principles of pertinence is then epiphenomenal—it is a consequence of the more basic features described by the principles of relevance.

Alternatively, the existence of the cognitive mechanisms described by RT may be denied. This would likely be the claim of most neo-Griceans: that (some version of) the cooperative principle is the fundamental cognitive feature that achieves the functionality described by the principles of pertinence. This view leads, however, to a conclusion that we may wish to resist. As already noted, natural selection tends toward mechanisms that can achieve the requisite functionality as precisely as possible (Fisher, 1930; Grafen, 2007; Hamilton, 1964). In the case of communication, the most precise mechanism would be some cognitive proxy of pertinence. The claim of this study is that this is precisely what relevance (as defined by RT) is: a proximate measure of pertinence. So if we deny the existence of such mechanisms, then we seem also to imply that the principles of pertinence *cannot* be cognitively implemented in such a direct manner—since if such a mechanism could be built then it would be. That it cannot be is a strong claim, whose foundations are opaque. Empirical verification would be necessary.

A second objection might be to point out that while the model assumes that both actors and reactors behave optimally, this is simply not the case with human communication. Misunderstandings occur. From an evolutionary perspective, however, this is not problematic. Behaviors need only be optimal on average in order to be selected. Many mechanisms in the natural world only approximate the functions for which they have been selected, for that is sufficient. Moreover, behavior in the natural world is frequently maladaptive, but the underlying mechanisms can still be selected for if they produce adaptive behavior on average—and this is all that is necessary for the model to work.

A third objection might be to observe that the model presented here depends on the fact that production and comprehension are sequential, yet empirical research on how humans actually behave in communication reveals that interlocutors' shared representations do not flow from one individual to another, but are rather negotiated upon in the course of conversation (Clark, 1996; Garrod & Anderson, 1987; Pickering & Garrod, 2004). However, this interactive approach to communication need not be juxtaposed with RT. Each individual

utterance, even in fast moving, overlapping dialog, may be performed according to the principles of relevance, and this may in turn give rise to shared representations. A relevance-theoretic analysis of the psycholinguistic data on dialog does not yet exist and may offer a fruitful avenue for future research.

The fourth and final objection would be to pursue the argument that RT is psychologically implausible, as relevance lacks any meaningful metric (Gazdar & Good, 1982). The point here is that relevance is a tradeoff between two phenomena, cognitive effects and processing effort, but there is no common metric by which we can measure the degree of either, and hence we cannot calculate relevance. The traditional response (Sperber & Wilson, 1995) has been to argue that relevance should be measured in comparative rather than absolute terms, for this is more likely to provide a psychologically plausible starting point. The evolutionary perspective developed in this study offers an additional way in which the problem of measurement may be overcome; namely that relevance is, ultimately, measured in terms of its contribution to the currency of natural selection: inclusive fitness. The basic idea is that if two individuals, identical in all ways until a particular moment, produce at that moment two different utterances, then the more relevant utterance will be the one that, on average, results in the greater inclusive fitness for the speaker. Such a metric does not lend itself to experimental investigation, of course, but it does offer a response to what is arguably the most substantial criticism of RT (Wedgwood, 2005).

5. Conclusion

The Gricean approach to communication, amended and refined in the light of subsequent developments, has dominated the field of linguistic pragmatics since its inception. RT offers a radical alternative that, if widely accepted, could be seen as a paradigm change in the field (Levinson, 1989). RT invokes evolutionary considerations as part of its justification for the first principle of relevance (Sperber & Wilson, 2002), but a formal examination of the evolutionary stability of communicative cooperation has not previously been developed. This study has shown that such an analysis speaks in favor of the proposed paradigm change: The functional roles of signals and responses in communication map quite directly onto the principles of relevance that lie at the center of RT. It is often forgotten that Grice's (1975) *Logic and conversation* was only intended to be tentative and programmatic, and that Grice himself recognized that a more secure foundation was necessary: ''I am… enough of a rationalist to want to find a basis that underlies these facts [of how people behave in communication] … I would like to think of the standard type of conversational practice not merely as something that all or most do *in fact* follow but as something that it is *reasonable* for us to follow, that we *should not* abandon'' (Grice, 1975, p. 48, italics in original). Natural selection can provide such a foundation, but it points toward RT, rather than to Grice's cooperative principle and its subsequent refinements.

Notes

1. The original wording is in fact ''Every *ostensive stimulus* carries a presumption of its own relevance'' (italics added). However, as our focus here is language, *ostensive stimulus* has been replaced by the more specific term *utterance*.
2. Named after German economist Heinrich Freiherr von Stackelberg, who studied duopolies in which one firm would act first and the other would then act knowing what the first firm had done (von Stackelberg, 1934).

Acknowledgments

The author was funded in this work by grants from the AHRC and ESRC. He wishes to thank Stu West, Dan Wedgwood, Tom Dickins, Andy Smith, Martin Pickering, Tamsin Saxton, and three anonymous reviewers for comments on earlier drafts, and Andy Gardner and Jim Hurford for helpful discussion of the ideas.

References

Atlas, J. D. (2005). *Logic, meaning and conversation: Semantic underdeterminacy*. Oxford, England: Oxford University Press.

Austin, J. L. (1955). *How to do things with words*. Oxford, England: Oxford University Press.

Barkow, J. H., Cosmides, L., & Tooby, J. (1992). *The adapted mind: Evolutionary psychology and the generation of culture*. Oxford, England: Oxford University Press.

Burling, R. (2005). *The talking ape: How language evolved*. Oxford, England: Oxford University Press.

Carston, R. (2002). *Thoughts and utterances: The pragmatics of explicit communication*. Oxford, England: Blackwell.

Clark, H. H. (1996). *Using language*. Cambridge, England: Cambridge University Press.

Dessalles, J.-L. (1998). Altruism, status and the origin of relevance. In J. R. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.), *Approaches to the evolution of language* (pp. 130–147). Cambridge, England: Cambridge University Press.

Diggle, S. P., Gardner, A., West, S. A., & Griffin, A. S. (2007). Evolutionary theory of bacterial quorum sensing: When is a signal not a signal? *Philosophical Transactions of the Royal Society B*, *362*(1483), 1241–1249.

Dunbar, R. I. M. (1997). *Grooming, gossip, and the evolution of language*. London: Faber.

Dunbar, R. I. M., & Barrett, L. (2007). *Oxford handbook of evolutionary psychology*. Oxford, England: Oxford University Press.

Fisher, R. A. (1930). *The genetical theory of natural selection*. Oxford, England: Clarendon.

Garrod, S., & Anderson, A. (1987). Saying what you mean in dialogue: A study in conceptual and semantic co-ordination. *Cognition*, *27*, 181–218.

Gazdar, G., & Good, D. (1982). On a notion of relevance: Comments on Sperber and Wilson's paper. In N. Smith (Ed.), *Mutual knowledge* (pp. 88–100). London: Academic Press.

Grafen, A. (2007). The formal Darwinism project: A mid-term report. *Journal of Evolutionary Biology*, *20*, 1243–1254.

Grice, H. P. (1971). Meaning. In D. Steinburg & L. Jakobovits (Eds.), *Semantics: An interdisciplinary reader* (pp. 53–59). Cambridge, England: Cambridge University Press.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics III: Speech acts* (pp. 41–58). New York: Academic Press.

Hamilton, W. D. (1964). The genetical evolution of social behavior. *Journal of Theoretical Biology*, *7*, 1–52.

Horn, L. R. (1984). Towards a new taxonomy for pragmatic inference: Q-based and R-based implicature. In D. Schiffrin (Ed.), *Meaning, form, and its use in context: Linguistic applications* (pp. 11–42). Washington, DC: Georgetown University Press.

Huang, Y. (2007). *Pragmatics*. Oxford, England: Oxford University Press.

Kinsella, A. R. (2009). *Language evolution and syntactic theory*. Cambridge, England: Cambridge University Press.

Kopytko, R. (1995). Against rationalistic pragmatics. *Journal of Pragmatics*, *23*, 475–491.

Leech, G. N. (1983). *Principles of pragmatics*. London: Longman.

Levinson, S. C. (1983). *Pragmatics*. Cambridge, England: Cambridge University Press.

Levinson, S. C. (1989). A review of relevance. *Journal of Linguistics*, *25*, 455–472.

Levinson, S. C. (2000). *Presumptive meanings*. Cambridge, MA: MIT Press.

Lycan, W. (2008). *The philosophy of language: A contemporary introduction*. New York: Routledge.

Maynard Smith, J., & Harper, D. G. C. (2003). *Animal signals*. Oxford, England: Oxford University Press.

Mayr, E. (1963). *Animal species and evolution*. Cambridge, MA: Harvard University Press.

Miller, G. F. (2000). *The mating mind*. London: BCA.

Noveck, I. A., & Sperber, D. (2004). *Experimental pragmatics*. London: Palgrave Macmillan Ltd.

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, *27*, 169–255.

Schiffer, S. (1972). *Meaning*. Oxford, England: Clarendon Press.

Scott-Phillips, T. C. (2007). The social evolution of language, and the language of social evolution. *Evolutionary Psychology*, *5*(4), 740–753.

Scott-Phillips, T. C. (2008). Defining biological communication. *Journal of Evolutionary Biology*, *21*(2), 387–395.

Scott-Phillips, T. C. (2010). Animal communication: Insights from linguistic pragmatics. *Animal Behaviour*, *79*(1), e1–e4.

Searcy, W. A., & Nowicki, S. (2007). *The evolution of animal communication*. Princeton, NJ: Princeton University Press.

Shannon, C. E., & Weaver, W. (1949). *The mathematical theory of information*. Urbana: University of Illinois Press.

Sperber, D., & Wilson, D. (1987). Précis of relevance: Communication and cognition. *Behavioral and Brain Sciences*, *10*(4), 697–754.

Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd. ed.). Oxford, England: Blackwell.

Sperber, D., & Wilson, D. (2002). Pragmatics, modularity and mind-reading. *Mind and Language*, *17*, 3–23.

von Stackelberg, H. (1934). *Marktform und Gleichwicht*. London: Springer Verlag.

Tinbergen, N. (1963). On the aims and methods of ethology. *Zeitschrift für Tierpsychologie*, *20*, 410–433.

Wedgwood, D. (2005). *Shifting the focus: From static structures to the dynamics of interpretation*. Oxford, England: Elsevier.

West, S. A., Griffin, A. S., & Gardner, A. (2007). Social semantics: Altruism, cooperation, mutualism and strong reciprocity. *Journal of Evolutionary Biology*, *20*, 415–432.

Wilson, D., & Sperber, D. (1981). On Grice's theory of conversation. In P. Werth (Ed.), *Conversation and discourse* (pp. 158–178). London: Croom Helm.

Wilson, D., & Sperber, D. (2002). Truthfulness and relevance. *Mind*, *111*, 583–632.

Wilson, D., & Sperber, D. (2004). Relevance Theory. In L. R. Horn & G. Ward (Eds.), *The handbook of pragmatics* (pp. 607–632). Oxford, England: Blackwell.

Yus Ramos, F. (1998). A decade of Relevance Theory. *Journal of Pragmatics*, *30*(3), 305–345.

Appendix

Consider a set of possible actions, A, and a set of possible reactions, R. Both are potentially infinite in size. For each pair of actions $a_i \in A$ and $r_j \in R$, there will be an associated pair of payoffs $\Pi_A(a_i,r_j)$ for the actor and $\Pi_R(a_i,r_j)$ for the reactor (relative to doing nothing). We define the null action a_0 and the null reaction r_0 such that $\forall i, j$, $\Pi_A(a_0,r_j) = \Pi_R(a_0,r_j) = \Pi_R(a_i,r_0) = 0$ and $\Pi_A(a_i,r_0) = -\varepsilon$, where ε reflects the opportunity and/or energy cost of the action. Call the actually performed actions and reactions a^* and r^*, respectively. It is axiomatic that participants are rational maximizers, so the actual payoffs achieved will be:

$$\Pi_R(a^*, r^*) = \max_j \Pi_R(a_i, r_j) \tag{1}$$

$$\Pi_A(a^*, r^*) = \max_i \Pi_A(a_i, r^*) \tag{2}$$

We want to show that if $a^* \neq a_0$, then $\Pi_R(a^*,r^*) > 0$, that is, that every non-null action anticipates a non-null reaction. This shows that a^* is pertinent, and once that is true then the fact that a^* is optimally pertinent follows immediately from eqn 2. To see that a^* is pertinent, suppose that $a^* \neq a_0$ but that, contrary to what we wish to show, $\Pi_R(a^*,r^*) \leq 0$. Then $\Pi_R(a^*,r^*)$ must be equal to 0, since $\Pi_R(a_i,r_0) = 0$. But if $\Pi_R(a^*,r^*) = 0$, then a^* can be equal to a_0, since $\Pi_R(a_0,r_j) = 0$. This contradicts our supposition, which therefore must be false, and hence our assertion that if $a^* \neq a_0$, then $\Pi_R(a^*,r^*) > 0$ must be true.