Being polite is a handicap: Towards a game theoretical analysis of polite linguistic behavior

Robert van Rooy*
Institute for Logic, Language, and Computation
University of Amsterdam
vanrooy@hum.uva.nl

Abstract

In this paper I argue for a broad game theoretical perspective on language use. Polite linguistic behavior, in particular, should be taken as rational interaction of conversational partners that each come with their own beliefs and preferences. I argue that the *function* of making a request in a polite way is to turn a situation in which preferences are not well aligned to one where they are by assuming that to utter polite expressions is *costly*. This idea will be formalized by making use of Lewisean *signaling games* and the biological *handicap principle*.

Using language rationally: broadening Gricean pragmatics

Within linguistic *pragmatics*, Grice's (1967) *cooperative principle* has always played an important role, the assumption that speakers are maximally efficient rational cooperative language users. Grice comes up with a list of four rules of thumb – the maxims of *quality*, *quantity*, *relevance* and *manner* – that specify what participants have to do in order to satisfy this principle. They should speak sincerely (*quality*), relevantly (*relevance*) and clearly (*manner*), and must provide sufficient information (*quantity*). Grice's maxims of conversation are sometimes seen as statements of regular patterns, or laws, of behavior. Problematic for such a view, however, is that the maxims are often violated in actual communication (see below). For some Griceans this can, almost by definition, never be the case: if the maxims appear to be violated, they are still obeyed – so it is argued –, though now at a 'deeper' level. However, such a move turns the down-to-earth Gricean rules of thumb not only into a mysteriously 'deep' theory, it also threatens to make the whole theory completely unfalsifiable and thus avoid of interest.

A more interesting perspective upon the Gricean principles, I believe, is to think of them as conventional *norms* evolved under the pressure of *evolution*: if agents behave in accordance with the norms, we are able to communicate *useful* information in an *efficient* but still *reliable* way. This suggests that formal analyses of evolution – in particular, evolutionary game theory (cf. Weibull, 1995) – could and should give

*Thanks to Manfred Krifka, who stimulated me most to write this paper, two anonymous TARK reviewers, Johan van Benthem, but especially Carl Bergstrom for insightful comments on an earlier version of this paper.

an explanatory account of how these norms/conventions could have emerged. A crucial advantage of taking the Gricean maxims as conventional *norms* – rather than as behavioral rules – of communicative behavior, is that now we don't have to 'explain away' cases in which the Gricean maxims are violated. Instead, we can now study the circumstances in which these violations occur.

As has been observed by Brown & Levinson (1978), Leech (1983), and others, Gricean maxims are systematically violated when *politeness* is of crucial importance. People pick 'safe topics' (e.g., the weather) to stress agreement and communicate an interest in maintaining good relations – but thereby violate the maxim of *relevance*. Euphemisms avoid mentioning the unmentionable, but in the process of using them people violate the maxims of *manner* and *quality*. Sentences typically lack their customary *quantity* implicatures (and thus violate the corresponding maxim to give as much relevant information as they can) in circumstances in which the speaker wishes to be modest, to avoid insulting the speaker, and so on.

1 Polite linguistic behavior: why and how?

A conversation is a multi-agent situation, and in such a situation each of its participants come with their own goals and preferences. Grice's (1967) influential cooperative principle – and the maxims that follow from it – implicitly assumes that in a conversation these goals and preferences are fully aligned. But they need not be, of course; they coincide only in special cases. Thinking of language as an instrument to influence other's behavior into your own advantage suggests that *strategic* considerations play a much more important role in language use than recognized by Grice.

A standard way in which we try to influence each other's behavior is by changing each other's beliefs by means of *assertions*. A more direct way of doing so, however, is by using *commands* and *requests*. In which circumstances is it rational to express what one wants the other to do in a polite, indirect, way? Our observation above that Grice's conversational maxims are typically violated in case politeness is at issue suggests a straightforward answer when we take a more general game-theoretical perspective: in these cases where we cannot assume the cooperative principle to hold, because the assumption behind this principle – the assumption that the goals and preferences of the agents are (known to be) perfectly aligned – is violated.

At first, it seems as if this hypothesis must simply be false: isn't it typically the case that for commands, speaker and hearer have opposing preferences? I don't believe so. First, in cases where there is no threat of imposition, as in instructions, one typically uses (short) imperatives: e.g. *Click here to see more!*. In fact, imperatives typically tend to be short, as if by convention. An evolutionary perspective suggests that only that part of the signal can become a communicative convention that is likely to influence the receiver in such a way that (on average) *both* the signaler and the receiver benefit from the exchange. But this suggests that this also holds for commands (see also Milikan, 1984). And indeed, pure and unexplained orders are sometimes used for the obvious benefit of the addressee:

(1) a. Enjoy yourself! Have fun!

b. Come in! Sit down! Don't worry about me!

In other cases, when imperatives are used as *warnings*, the advantage for the addressee can't be 'read off' immediately from the linguistic expression, but is still really there:

(2) a. Be careful! He's a dangerous man.

b. Duck!

For (2b), for instance, it might be better for me to rapidly follow the 'master's' instruction than to respond more slowly (too late) after accessing the situation and deciding for myself on the proper action.

Thus, I would claim that for direct commands the (overall) preferences of speaker and hearer are not really different. (Sometimes the speaker can simply afford herself to be direct, because she is mutually known to have enough power to hurt the addressee if he doesn't comply to the order. But this makes obeying the order in the interest of the addressee as well!) At other times, however, we seek to influence others in order to let them do things that go against their immediate goals without (using) these extra powers. In these cases we typically use *polite requests*. Making a request by means of an indirect *polite* question *Could you by any chance lend me your car?* rather than in a directly commanding tone *Lend me your car!* is a typical case. But this suggests that the function of being polite is to *change* the original situation, or game, where the preferences of the conversational participants are not aligned to one where they are. In the latter situation, coordination (i.e. help of the other participant) can again be expected. A *first strategy* is to *pretend as if* preferences are already (more or less) aligned, or goals are shared. A first way of doing so is to use inclusive 'we':[1]

(3) a. Let's stop for a coffee. (i.e. *I* want a coffee, so let's stop)

b. Let's get on with diner, eh? (i.e. *you* should get on)

Another is to use linguistic markers, like *so* or *then*, which indicates that the speaker is drawing a conclusion to a line of reasoning carried out cooperatively with the addressee. They can be used as a *fake* prior agreement to pressure the addressee to accept the request/offer:

(4) a. So, I'll be seeing you at 5 o'clock then.

b. Take this radio off my hands for 5 quid then?

A third is to pretend to be great friends, lovers or members of a close group (by using slang):

(5) a. Help me with this bag here, will you pal/luv?

b. Lend us two *bucks* then, wouldja Mac?

If an already existing alignment of preferences cannot seriously be pretended, one can try a *second strategy*, to *minimize* the imposition/request as in (6):

(6) I *just* want to ask you if I can borrow a *tiny bit of* paper.

[1]Most examples below come from Brown & Levinson (1978), although they categorize them in a different way.

If even that doesn't work, things get serious and one has to use a *third strategy*. In these cases one can try to influence the other by (i) incurring a *debt* to one's conversational partner if she performed the requested action, or (ii) by reducing one's social status. The former can either be done directly, as in (7a), or more indirectly, as in (7b) and (7c), by signaling a *negative expectation*. In both cases the speaker signals that afterwards it is common ground that she owes her conversational participant something.

(7) a. I'll never be able to repay you if you ...

b. I don't suppose there'd be any possibility of you ...

c. You don't have any manila envelopes, do you by any chance?

The latter way of influencing the other, i.e. by reducing one's social status, is already done by phrasing what one wants the other to do in terms of a question (request) rather than as a direct command. Additionally, one can explicitly understate, or play down, one's self and one's capacities, as in (8a) to ask for help, or one can show hesitation (merging reluctance and incompetence), typically done in English by using 'uh' as in (8b), and (according to Brown & Levinson, 1978) in Tzeltal by using high pitch:

(8) a. I think I must be absolutely stupid but I simply can't understand this map.

b. I think you should, uh, attend to your files.

If you want someone else to do something that is not in accordance with her immediate goals or desires, you can try to influence her by promising her money: a transferable good. One can think of such a promise, e.g. (9), as a *contract*. I think that you will perform the for yourself undesirable action only if I pay for it. It will *cost* me money.

(9) If you mow the lawn, I'll give you five euros.

Just like this explicit promise involves a (proposal for) transfer of goods, so does, I propose, a polite request. Both turn the status quo situation of partial conflict into one where both parties are better off. However, in case of a polite request the transfer involves incurring a social debt and reducing one's social status, which are not material goods. But that doesn't matter. The essential point is that both are disadvantageous, and thus *costly*, for speakers.

It is a general rule within linguistics that an (un)marked expression tends to have an (un)marked meaning. As we will see in the following section, the markedness of an expression is typically 'measured' in terms of its complexity, and that of a meaning in terms of its probability. In this section, however, we saw that polite requests are marked compared to direct commands in terms of the *costs* that they incur. This, I will propose, has also a consequence for what one should think in these cases as a marked meaning. In the following more formal sections I will propose that the way we should think of 'marked meanings' can be characterized in terms of the way the goals and preferences of conversational partners are aligned. In case the preferences are strongly aligned, a marked expression suggests an *unlikely* meaning, at other times it suggests that the speaker is a *high quality* signaler. I will formalize this by making use of Lewisean (1969) *signaling games* and Zahavi's (1975) biological *handicap principle* (and/or Spence's analysis of education).

2 Signaling games

2.1 Costless signaling

A signaling game is a game of incomplete information where one individual, the signaller, is informed of the value of an uncertain parameter t (her type) and then chooses an action m, referred to as a message. A second individual, the receiver, observes this message (but not the value of t) and performs some action e. The payoff to each individual depends only on the value of t and the action e adopted by the receiver.

For simplicity I will assume that the strategies of sender and receiver are *functions* from types to messages, and from messages to actions, respectively. Assuming that T is the set of types, M the set of messages, and E the set of actions, a sender-strategy S is thus an element of $[T \to M]$ and a hearer-strategy R is an element of $[M \to E]$.

Although the messages used in signaling games need not have a pre-existing meaning, they can acquire one due to the strategic interaction between sender and receiver, in particular, when they use strategies that are part of a *Nash equilibrium*. A strategy profile $\langle S, R\rangle$ forms a Nash equilibrium iff neither the sender nor the receiver can do better by unilateral deviation. Signaling games as described above above typically have lots of equilibria (depending on probability and payoff functions). However, we are interested only in equilibria of a particular kind: equilibria where different types of senders send different messages. These equilibria are called (totally or partially) *separating*. If a signaling game has such a separating equilibrium, the sender can reveal some information about her type. In other words, in such a situation communication is possible. If sender strategy S is part of a separating equilibrium, we might say that message $S(t)$ *means* $\{t' \in T : S(t') = S(t)\} = S_t$, because all and only individuals of the types in S_t send message $S(t)$. Notice, however, that in a two-type two-message situation there might already be 2 separating equilibria. For this reason, one concentrates in costless signaling games not so much on one particular equilibrium and what the particular messages mean in this equilibrium, but rather on whether separating equilibria exist. If there does not exist a separating equilibrium in a game with two-types, no communication is possible.

For the simple situation with two types of senders and where the receiver can choose between two actions, we can show that separating equilibria exist in very specific situations only (see Gibbons, 1992). Consider the following abstract table:

two-type, two-action:

	e_H	e_L
t_H	$x, 2$	$z, 0$
t_L	$y, 0$	$w, 1$

It is easy to see that in this two-type, two-action situation, communication (i.e. a separating equilibrium) is possible only in case $x \geq z$ and $y \leq w$. We can check this by looking at the other possible cases: (i) if $z > x$ and $y > w$ the preferences are strictly opposed. No communication is possible now, because t_H would like the hearer to believe that his type is t_L, and the other way around for a sender of type t_L. Thus the signaling game will have no Nash equilibrium; (ii) if $x > z$ and $y > w$ (or if $z > x$ and $w > y$) *both* types of sender prefer the same action of the receiver: in our example both prefer action e_H to action e_L. Also in this case no communication will take place, because both players want the receiver to believe that her type is t_H. We can conclude that in the simplest two-type, two-action situations, communication is possible only in case

the preferences are perfectly aligned. In an important paper, Crawford & Sobel (1982) have generalized this simple result: they show that the amount of possible (credible) communication in costless signaling games depends on how far the preferences of the agents are aligned.[2]

2.2 Costly signaling and efficient language organization

Above we only considered signaling games with costless messages. In standard game theory, however, the messages used can be more or less expensive. These extra costs of signals can influence the equilibrium play of the game. But this means that these costs can have an effect on the meaning of the signals as well. In a signaling game with payoff relevant messages, the payoff-functions of the sender (U_1) and the receiver (U_2) are elements of $[T \times M \times E \rightarrow \mathbf{R}]$. A strategy profile $\langle S, R\rangle$ together with probability function P forms a (sequential) *Nash equilibrium* iff neither the sender nor the receiver can do better by unilateral deviation. That is, $\langle S, R\rangle$ forms a Nash equilibrium iff for all $t \in T$ the following two conditions are obeyed:

$$\begin{array}{ll} \text{(i)} & \neg\exists S' : U_1(t, S(t), R(S(t))) < U_1(t, S'(t), R(S'(t))) \\ \text{(ii)} & \neg\exists R' : \sum_{t' \in S_t} P(t'/S_t) \times U_2(t', S(t'), R(S(t'))) < \\ & \qquad \sum_{t' \in S_t} P(t'/S_t) \times U_2(t', S(t'), R'(S(t'))) \end{array}$$

If the participants of a conversation coordinate on strategy profile $\langle S, R\rangle$, the meaning of the message $S(t)$ is S_t, and depends partly on the costs of messages. Suppose that we can measure the cost of a message $S(t)$ by its length, $l(S(t))$. Suppose also that the preferences of sender and receiver are perfectly aligned and that successful communication is most important. To account for the latter, let us assume temporary that $E = T$ and that a sender of type t can communicate his type successfully when communication strategy combination $\langle S, R\rangle$ is used iff $R(S(t)) = t$. Taking also the costs of messages into account, the utility of a type-message-action triple can then naturally be defined in terms of the success of communication and the length of the message used as follows:

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

Even with extra costs, such games of complete coordination still have many equilibria. However, the expected utilities of these equilibria differ. The expected utility of equilibrium $\langle S, R\rangle$ with respect to probability function P is simply $\sum_{t \in T} P(t) \times U(t, S(t), R(S(t)))$. It is obviously the case that *pooling* (i.e. non-completely separating) equilibria will have a lower utility than completely separating ones. However, by making utilities depending on the length of the messages we can also distinguish between separating equilibria where we have full communication. Equilibria with a higher utility are those that – on average – send messages with a shorter length.

Standard game theory does not really distinguish equilibria with respect to expected utility, and neither does the standard (ESS) solution concept in *evolutionary* game theory (because all separating Nash equilibria are *strict* ones). However, when we either give up the assumption that the evolutionary process behind the ESS-concept makes

[2]For more discussion on credible communication in costless signalling games, see Van Rooy (2003).

use of *random pairing* (e.g. Hamilton 1964; Skyrms, 1996) or is *deterministic* (Kandori, Mailath & Rob, 1993; Young, 1993) (in which case phonological simplification can be due to *mutation*), we can show (van Rooy, in press) that in pure coordination games evolution tends to select equilibria with, on average, messages with small(est) length.[3] This explains the fact observed by Zipf (1949) and others that messages that are used a lot (have a probable meaning) tend to be short: it is an evolutionary explanation behind the economics of language organization.

The above reasoning suggests that costly messages are typically used to express meanings with a low probability. In section 1 we argued that polite expressions are typically more costly than their less polite counterparts. However, the reason behind this doesn't seem to have to do a lot with the probability of the meaning expressed. To explain the reason why polite expressions are costly we have to look somewhere else. In these cases, the preferences of the conversational partners are not in perfect harmony. In the final sections of this paper I will propose that the explanation for the use of costly polite expressions is given by the biological handicap principle.

3 Politeness: signaling a handicap to establish harmony

3.1 The handicap principle

As we have seen in section 2.1, in two-type, two-action costless signaling games the sending of signals is useless unless the preferences of the players are in perfect harmony. This is in accordance with Maynard Smith's (1974) and Dawkins & Krebs' (1978) early use of game theory to account for animal 'communicative' behavior. They conclude that animals with (partially) conflicting interests will not communicate: threat display conveying accurate information about aggressiveness or level of escalation is not evolutionarily stable, and the animals will maintain a 'poker face' to hide their true intentions. But even in these simple situations, agents – animate or human – sometimes send messages to each other, even if the preferences are less harmonically aligned. Why would they do that? Moreover, could there somehow still be real honest communication going on? If so, how?

In the animal kingdom we see traits which are truly exagerating: peacocks with very long tails, stags with enormous antlers, etc. Such exaggerated traits are not only costly in terms of their production and maintenance, but are also deterious with regard to survival. Natural selection should therefore eliminate the showy males and favor the more cryptic ones. How can it be that natural selection didn't do so? Zahavi (1975) proposed an appealing explanation: individuals who sport the exaggerated traits and live to tell the tale must be truly extraordinary males, with genotypes that can readily tolerate the survival costs of the trait. Consequently, Zahavi argued, females should pick males with these 'handicaps' because they have made it through a survival filter. So, by showing one's handicap, an agent can communicate his true quality/ability in an honest way. Handicaps make honest communication possible, even if the preferences of the individuals involved are not fully aligned.[4]

[3] It is important that we talk about pure coordination games, because only in those cases it is guaranteed that the *risk-dominant* strategy pairs as selected by the Kandori, Mailath & Rob-approach are also the Pareto efficient ones.

[4] The economist and sociologist Thorstein Veblen (1989) already suggested a similar explanation for the seemingly ridiculous squandering of resources by the wealthy classes he observed: The wealthy

Over the years, the use of this *handicap principle* has been extended from sexual selection to a number of other animal communicative behaviors. Noteworthy are the analyses of begging baby birds, alarm calls, and threat behavior of animals when contesting resources (see Bergstrom's http://octavia.zoology.washington.edu/handicap/). These analyses assume that there exists a variety of ways in which communicative displays can handicap the signaler. That is, in different signaling systems, the costs of messages are determined in different ways. In the original examples of sexual signaling by peacocks and others, the costs were merely *costs of production*. For threat display – when signals are sent for a *strategic* reason –, however, the costs are not incurred by production (these are negligible), but by the receiver. The receiver can verify the message (by ignoring the threat, i.e. attack), and the cost depends on the type of the sender: high quality senders do better (when attacked) than low quality senders. In these cases, the costs might be called *social costs*.

3.2 A game theoretical analysis

Zahavi stated his handicap principle informally. However, Grafen (1990) showed that it can be formalized by making use of game theory (and he gives an extra motivation by thinking of it from an *evolutionary* point of view). As it turns out, Grafen's analysis is essentially the same as Spence's (1973) model of job-market signaling, making use of Lewisean (1969) *signaling games* with added *costs*.[5] The utility functions of sender and receiver now depend not only on the sender's type and the receiver's response – as in costless signaling games –, but also on the message sent. We have seen above that making messages more or less costly makes much sense in biological applications. The length of a peacock's tail, for example, is thought of as a signal and sending the 'I have an enormous tail'-signal costs a lot of extra energy – is a handicap – during life. While this biological signal is genetically determined, others – such as threat behavior – are more strategic in nature. The assumption that signals can be expensive makes much sense in economical applications as well: Education, advertisements, and pricing habits, for instance, are thought of as signals, and the sending of these signals can be more or less *costly* than others. Notice that in economical applications sending an expensive message is almost always done for a strategic reason.

Consider the abstract two-type two-action game of section 2.1 again with $x = 1 > 0 = z$ and $y = 1 > 0 = w$. This is a favorite type of signaling game studied by economists and biologists because it is the simplest kind of game in which the agents' preferences are neither perfectly aligned nor strictly opposed, and where the role player prefers, irrespective of her type, column player to choose e_H.

For a separating equilibrium to exist, individuals of type t_L must not benefit by adopting the signal typical of individuals of type t_H, even if they would elicit a more favorable response by doing so. The solution suggested by Spence (1973) and Zahavi (1975) is to make the signal typical of individuals of type t_H costly to produce, particularly for individuals of type t_L. Assume that m' is cost-free, but that the cost of m depends on the type of the signaler. The cost is denoted $C(t_H, m)$ for individuals of type t_H, and $C(t_L, m)$ for individuals of type t_L. Provided that $C(t_L, m) > 1 > C(t_H, m)$,

engage in conspicuous consumption in order to advertise, i.e. *signal*, their wealth.

[5] Spence's analysis has become a textbook classic within economics, especially since the increasing recognition of the importance of (asymmetric) information for economic reasoning (cf. Hirschleifer & Riley, 1992).

the cost of m will outweigh the benefits of its production for individuals of type t_L, but not for individuals of type t_H, so that the following seperating equilibrium exists: individuals of type t_H send message m, while individuals of type t_L send (if at all) message m'. Alternatively, we can reach the same conclusion by assuming that $x > C(t_H, m) = C(t_L, m) > y$.[6]

To make things concrete, let us assume that we have two types of individuals, two messages, and two actions that the receiver can perform. On our simplifying assumption that sender and receiver strategies are *functions*, a sender strategy, for instance, is a function from types to messages: we don't allow for mixed sender strategies. This means that both sender and receiver have four possible strategies:

Sender :

	t_H	t_L
S_1	m	m'
S_2	m	m
S_3	m'	m
S_4	m'	m'

Receiver :

	m	m'
R_1	e	e'
R_2	e	e
R_3	e'	e
R_4	e'	e'

Assume that senders of type t_H are less likely than senders of type t_L, e.g. $P(t_H) = \frac{1}{4}$. In that case it follows (by Bayes' law) that with respect to S_2 and S_4 – where both types of senders send the same message – $P(t_H/m) = P(t_H/m') = \frac{1}{4}$. We also make the crucial assumption of Spence and Zahavi that the cost of signal m is negatively correlated with the sender's quality. For simplicity we take t_H's cost of sending message m to be $\frac{1}{2}$, $C(t_H, m) = \frac{1}{2}$, while $C(t_L, m) = \frac{3}{2}$. We assume also the following situation in the 'beginning state' (with costless signal m') and the situation when m is sent:

m' :

	e	e'
t_H	$1, 2$	$0, 1$
t_L	$1, 0$	$0, 1$

m :

	e	e'
t_H	$\frac{1}{2}, 2$	$-\frac{1}{2}, 1$
t_L	$-\frac{1}{2}, 0$	$-\frac{3}{2}, 1$

The utilities in the latter game are the same as the benefits in the former, except that the utilities for the sender of a certain type are reduced by the cost of the message for individuals of that type. Notice that if m' is sent (normally interpreted as saying nothing), the hearer will choose e' because this action gives her a higher expected utility than e: 1 versus $(\frac{1}{4} \times 2) + (\frac{3}{4} \times 0) = \frac{1}{2}$. By sending m, however, the speaker can change the game. With the benefits, costs, and probabilities in place, we can determine the payoffs of each sender-receiver combination for each type:

[6] A more general characterization can be given. The utility of a signaler of type i to send message m_j if the receiver performed action e_k, $U_s(t_i, m_j, e_k)$, can be decomposed in terms of benefits and costs: $U_s(t_i, m_j, e_k) = B_s(t_i, e_k) - C(t_i, m_j)$, while the utility for the receiver, U_r, depends only on the type of the signaler and the action performed and thus equals the benefit. The benefits are such that all signalers strictly prefer e_H to e_L, while the receivers want correlation: $B_r(t_H, e_H) > B_r(t_H, e_L)$ and $B_r(t_L, e_L) > B_r(t_L, e_H)$. The value of the receiver's action e_H for a signaler of type t_i, V_i, can be defined as follows: $V_H = B(t_H, e_H) - B(t_H, e_L) > 0$ and $V_L = B(t_L, e_H) - B(t_L, e_L) > 0$. The cost of signaling m for a signaler of type i, C_i, is $C_H = C(t_H, m) - C(t_H, m')$ and $C_L = C(t_L, m) - C(t_L, m')$. In order for the desired separating equilibrium to exist, it must be the case that $U_s(t_H, m, e_H) > U_s(t_H, m', e_L)$ and $U_s(t_L, m', e_L) > U_s(t_L, m, e_H)$. This means that the following two conditions have to be fulfilled: $B(t_H, e_H) - C(t_H, m) > B(t_H, e_L) - C(t_H, m')$ and $B(t_L, e_L) - C(t_L, m') > B(t_L, e_H) - C(t_L, m)$. But in the situation under discussion this is the case exactly if $V_H < C_H$ and $V_L < C_L$. The conditions stated in the main text, i.e. $C(t_L, m) > (x - z) = (y - w) > C(t_H, m)$ and $(x - z) > C(t_H, m) = C(t_L, m) > (y - w)$, are both special cases of this.

t_H:

t_H	R_1	R_2	R_3	R_4
S_1	$\boxed{\frac{1}{2}, 2}$	$\frac{1}{2}$,2	$-\frac{1}{2}, 1$	$-\frac{1}{2}, 1$
S_2	$\frac{1}{2}, \frac{1}{2}$	$\frac{1}{2}, \frac{1}{2}$	$-\frac{1}{2}, 1$	$-\frac{1}{2}, 1$
S_3	0,1	$\boxed{1,2}$	$\boxed{1,2}$	0,1
S_4	0,1	$1, \frac{1}{2}$	$1, \frac{1}{2}$	$\boxed{0,1}$

t_L:

t_L	R_1	R_2	R_3	R_4
S_1	$\boxed{0, 1}$	1,0	1,0	$\boxed{0,1}$
S_2	$-\frac{1}{2}, \frac{1}{2}$	$-\frac{1}{2}, \frac{1}{2}$	$-\frac{3}{2}, 1$	$-\frac{3}{2}, 1$
S_3	$-\frac{1}{2}$,0	$-\frac{1}{2}$,0	$-\frac{3}{2}$,1	$-\frac{3}{2}$,1
S_4	$\boxed{0,1}$	$1, \frac{1}{2}$	$1, \frac{1}{2}$	$\boxed{0,1}$

For $\langle S, R\rangle$ to be an equilibrium in a signalling game, it has to be a Nash equilibrium for all possible types, i.e., both for t_H as for t_L. This means that the complete game has two equilibria: $\langle S_1, R_1\rangle$ and $\langle S_4, R_4\rangle$. (As the reader might check for herself, exactly the same reasoning goes through when we assume that $C(t_H, m) = C(t_L, m)$, but $U_1(t_H, e) > U_1(t_L, e)$.) This is enough to show that under natural assumptions a separeting equilibrium, i.e. $\langle S_1, R_1\rangle$, exists where a costly message is sent if and only if the sender is of a high type. Notice that because in equilibrium $\langle S_4, R_4\rangle$ the costly message m will never be sent, the fact *that* the sender uttered m already suggests that he is assuming the other, seperating, equilibrium.[7]

3.3 Being polite is a handicap

Notice that by using a *costly* message, we make a separating equilibrium possible *because* we have *changed the situation* from one where the preferences were not aligned to one where they are. I would like to propose now that there exists a correspondence between, on the one hand, the use of *costly* signals to insure *honest* communication in the animal kingdom, and, on the other, the use of *polite* linguistic expressions in human communication to signal *good intentions*.

Think of the original situation (where the sender signals m') as one where the receiver (player 2) wonders whether she should perform e or not (in which case she does nothing, or e'). Player 2 knows that the sender prefers her to take action e. However, player 2 doesn't know whether the sender is a grateful individual (of type t_H) who will reward her afterwards, or not (type t_L). In other words, the game that is being played is one of *incomplete information*. Let us say that $U_2(t_H, e) = 2$, but $U_2(t_L, e) = 0$. The receiver would like to perform e for a grateful individual, but otherwise prefers to do nothing (play e'). Let's say that $U_2(t_H, e') = U_2(t_L, e') = 1$. The game as described above shows that in case the sender can send a costly request (message) m to the receiver to do e such that $C(t_H, m) = \frac{1}{2}$ and $C(t_L, m) = \frac{3}{2}$, it is possible for a high, but not for a low, quality individual to influence the receiver to perform the desired action e. But how can it be that in this case the cost of the request is higher for one type of individual than for the other? In fact, how should we interpret the cost of the message in the first place?

To explain polite linguistic behavior in terms of costly messages, it seems clear that production costs shouldn't really play a role (in that sense, talk is cheap). This means

[7] Note that because the average utility of equilibrium $\langle S_1, R_1\rangle$ is higher than that of $\langle S_4, R_4\rangle$, one can also imagine that it will be singled out as the *unique* equilibrium according to an (even) more fine-grained equilibrium concept. To my surprise, biologists are normally already content to see that a separating equilibrium is possible, without wondering themselves under which circumstances this equilibrium will actually be selected.

that if the message is costly, this must partly be the receiver's responsibility. Thus, it are social costs that are at issue here. But what could these costs be? In section 1 we argued that there are (at least) two ways in which one can try to influence one's conversational partner to do something that goes against her own interest if one doesn't have enough power to force her, or cash money to pay: one can either (i) reduce one's social status or (ii) incur a social debt (with respect) to one's conversational partner. Making a polite request can be costly in both of these ways. However, some types of individuals can afford to pay a higher price (especially, incur a greater debt) than others: when asked for, for instance, some can pay the debt, i.e., do a favor of the same magnitude in return, while others cannot. This latter fact assures that (if paying the cost can be assured) only individuals of a high quality can afford to be polite.

3.4 Politeness and complexity

In the beginning of the paper we observed that if politeness plays an important role in communication, Griceans maxims tend to be violated. As it turns out, this is in particular the case for the maxim of *manner*: the rule that we should communicate our beliefs and desires in an *efficient* way. Indeed, whereas an economical view on language (use) suggests a preference for simple linguistic forms (see section 2.2), a pressure for being polite typically leads to complication rather than simplification of the form-meaning correspondence. Polite expressions are typically less direct than their less polite counterparts, and less direct linguistic expressions are typically more complicated than more direct ones. Compare:

(10) a. Come here!

b. Could you come here for a moment, please?

(11) a. I want to see you for a moment.

b. I wondered if I could possibly see you for a moment.

Above we have interpreted polite requests as being costly and suggested why it thus (normally) is a reliable signal of intentions.[8] Whether a sentence is used as a command or a request, however, can (normally) not be determined solely by the sentence's syntactic features. Often, one recognizes what people are doing in verbal interchanges (e.g. requesting, offering, criticizing, complaining, suggesting) not so much by what they overtly claim to be doing as in the fine linguistic detail of their utterances. Clues, like *phonological realization* and *length* of the sentence, are crucial here. So why do we associate the *amount of effort* used to make a request/command with the *force* with which the request/command is made? Why do polite requests tend to be more complex (longer) than direct commands?

Our approach in terms of costly signaling does not yet give an explanatory analysis of why a polite use of language leads to complication: the complexity of an utterance by itself cannot be costly enough (in terms of production costs) to induce the favored reaction by a Zahavian handicap reasoning. In this sense talk is pretty cheap. But why then this complexity? The reason is twofold, or so I would like to propose. The most important reason for complexity is simply that due to an extra expression of

[8] Unfortunately, the social cost mechanism isn't perfect. Slimy fellows can survive.

unexpectedness or gratefulness the social costs increase, and the individual that sends the message thus becomes more reliable. But the extra complexity is also used as an extra *marker* of politeness: a *third party* will notice that a polite request is being made which makes the reduction of the speaker's social status more serious, and gives more 'insurance' that the incurred debt indeed will be payed.

4 Conclusion and outlook

In this paper we have observed that polite speech typically violates the Gricean picture of language use as an efficient mechanism of transferring reliable and relevant information. I have argued that the reason behind this is that polite linguistic behavior can be expected when the crucial assumption behind Grice's cooperative principle does not hold: in case there is a conflict between the preferences of speaker and addressee. Being polite is seen as a strategic way of getting what one wants, and analyzed in terms of costly signaling games and Zahavi's handicap principle. Polite utterances come with social costs that can establish a harmony of preferences between sender and receiver that did not exist before, but these costs can be afforded only by certain types of individuals.

In this paper I have thought of polite speech from an economical cost-benefit point of view. However, factors not mentioned until now play a role as well, and it is not completely clear how they fit into the picture. For instance, the primary reason why the following sentences are progressively more polite seems to be related with the increasing choice left to the hearer to opt out (cf. Leech, 1983):

(12) a. Take me home!

b. Can you take me home?

c. *Could* you take me home?

d. Could you *possibly* take me home?

Although the hearer's opting out is costly for the speaker, it is not obvious how to model this increase of politeness in terms of costly signaling, in particular, in terms of the handicap principle. Another limitation is that I concentrated myself here only on commands and requests. However, politeness plays a crucial role in other speech acts as well. For offers, for instance, it is more polite to offer more than offering less, and by doing so in a commanding tone leaving no other options: *Take some cookies!* is better than *You may take a cookie.* An assertion, as another example, can be polite if it uses, for instance, in-group identity markers, or if it shows (or exagerates) an interest, approval, sympathy, or respect for the hearer (e.g. by using honorifics). Intuitively, the handicap analysis could be of help here as well, because those ways of being polite can be costly too. This is obvious for the above offers, and the polite assertions come with the danger that the addressee will take the speaker too seriously. In general, however, it remains unclear to me in how far Brown & Levinson's (1978) essentially *hearer* oriented 'preservation of *face*' analysis of politeness can be captured in terms of our essentially *speaker* oriented costly-signaling account. To give more weight to the hearer's benefits, the game-theoretical analysis of *bargaining* should perhaps be relevant here as well. I hope to get clearer about this in the future.

The handicap principle was introduced by Zahavi to account for *honest* communication. As far as *human* language use is concerned, this suggests that it should be used primarily to motivate Grice's *maxim of quality*. This maxim asks of the speaker (of an assertion) not to say more than he has evidence for. Indeed, in an interesting recent paper Lachman et al. (2001) suggest that by making use of social costs, the handicap principle can be used here. Why should a speaker obey Grice's maxim? Sometimes it is advantageous pretending to know more than one actually does, or even to lie. In normal conversations, however, if one makes a statement, one is also *committed* to its truth. The truth of this statement can be *verifiable*. Thus, to make a strong statement can be *costly*: you can be punished (perhaps in terms of reputation) when you have claimed something that turns out not to be true. And this can be enough not to violate the quality maxim.

Both my analysis of polite requests as the above sketched analysis of truthful human communication suggests that although natural language expressions are cheap in production, the handicap principle can still be used to account for communicative behavior between humans. With Lachman et al (2001) I take this to be an important insight: it suggests a way to overcome the limitations of both cheap talk signaling and Gricean pragmatics. The handicap principle shows us how to account for successful communication even if the preferences of the agents involved are not well aligned.

References

[1] Brown, P. and Levinson, S.C. (1978). 'Universals in language usage: politeness phenomena'. In E. Goody (ed.), *Questions and politeness: strategies in social interaction*, pp. 56-311. Cambridge Papers in Social Anthropology 8. Cambridge University Press.

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

[3] Dawkins, R. and J. Krebs (1978), 'Animal signals: information or manipulation', In: J. Krebs & N. Davies (eds.), *Behavioural Ecology: an Evolutionary Approach*, Blackwell Scientific Publications, Oxford, 282-309.

[4] Gibbons, R. (1992), *A primer in Game Theory*, Harvester Wheatsheaf, New York.

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

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

[7] Hamilton, W.D. (1964), 'The genetic evolution of social behavior', *Journal of Theoretical Biology*, **7**: 1-52.

[8] Hirshleifer J. and J.G. Riley, (1992), *The Analytics of Uncertainty and Information*, Cambridge surveys of Economic Literature, Cambridge University Press, Cambridge, Massachusetts.

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

[10] Lachman, M., Sz. Szamado, and C. Bergstrom (2001), 'Cost and conflict in animal signals and human language', *Proceedings of the National Academy of Sciences USA*, **98**: 13189-13194.

[11] Leech, G. N. (1983), *Principles of Pragmatics*, Longman, London.

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

[13] Maynard Smith, J. (1974), 'The theory of games and the evolution of animal conflict', *Journal of Theoretical Biology*, **47**: 209-221.

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

[15] Rooy, R. van (2003), 'Quality and quantity of information exchange', *Journal of Logic, Language and Information*, **12**, vol 6, to appear.

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

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

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

[19] Veblen, T. (1899), *The Theory of the Leisure Class*, Dover Publications, New York.

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

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

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

[23] Zipf, G. (1949), *Human behavior and the principle of least effort*, Cambridge: Addison-wesley.