Politeness and reputation in cultural evolution

Roland Mühlenbernd^{1,2} · Sławomir Wacewicz¹ · Przemysław Żywicznyński¹

Accepted: 9 October 2020 / Published online: 22 December 2020

© The Author(s) 2020

Abstract

Politeness in conversation is a fascinating aspect of human interaction that directly interfaces language use and human social behavior more generally. We show how game theory, as a higher-order theory of behavior, can provide the tools to understand and model polite behavior. The recently proposed responsibility exchange theory (Chaudhry and Loewenstein in *Psychol Rev* 126(3):313–344, 2019) describes how the polite communications of thanking and apologizing impact two different types of an agent's social image: (perceived) warmth and (perceived) competence. Here, we extend this approach in several ways, most importantly by adding a cultural-evolutionary dynamics that makes it possible to investigate the evolutionary stability of politeness strategies. Our analysis shows that in a society of agents who value status-related traits (such as competence) over reciprocity-related traits (such as warmth), both the less and the more polite strategies are maintained in cycles of cultural-evolutionary change.

Keywords Politeness · Reputation · Cultural evolution · Responsibility exchange theory · Evolutionary game theory

1 Introduction

Why do we say ‘I am sorry’ or ‘thank you’? Or conversely, why don’t we do this all the time? Taking inspiration from a long tradition of research into strategic use of language (cf. Grice 1975; Brown and Levinson 1978; Pinker et al. 2008; Lee and Pinker 2010; Franke and Jäger 2016), we see many forms of polite language use, such as apologies or expressions of gratitude, as instances of *strategic use of conversational politeness* (SCP). We mean ‘polite’ in a basic intuitive sense in which competent speakers intuitively recognize certain linguistic communications as polite (Leech 1983; Eelen 2001; Watts 2003; Żywicznyński 2012; Brown 2017), and ‘strategic’ in an economic sense in

✉ Sławomir Wacewicz
wacewicz@umk.pl

¹ Nicolaus Copernicus University, Toruń, ul. W. Bojarskiego 1, 87-100 Toruń, Poland

² Leibniz-Zentrum Allgemeine Sprachwissenschaft, Berlin, Germany

which rational agents choose strategies that maximize their utilities. Virtually all existing utility-based accounts of SCP (Pinker 2007; Clark 2012; Asher and Quinley 2012; Quinley 2012; Quinley and Ahern 2012) assume the costs of polite communications to be cashed out in the social commodity ‘face’ (Goffman 1967; Brown and Levinson 1987; Culpeper 2011; Brown 2017), without however a satisfying account of the benefits of SCP, nor the mechanism through which face acquires its purchasing power. In contrast, the recently proposed ‘Responsibility Exchange Theory’ (Chaudhry and Loewenstein 2019) explains SCP in a class of situations involving a transfer of credit or blame by grounding both the costs and benefits in somewhat more tangible social constructs: (perceived) competence and (perceived) warmth.

Responsibility exchange theory constitutes a marked advance in the understanding of SCP, but remains confined: firstly, to narrow contexts, and secondly, to an epistemic model, which explains and predicts individual choice of more versus less polite strategies (thanking, apologizing versus bragging, blaming). Building on the insights from Chaudhry and Loewenstein (2019), C&L hereafter, here we take further steps in extending their epistemic model into a more general evolutionary model of SCP. Rather than modeling the utility of individuals in one-off dyadic interactions, we model the fitness of behavioral strategies that compete with one another over time and inside a population; in short, strategies of politeness are subjected to the dynamics of cultural evolution. We show that in a society of agents who value status-related traits (such as perceived competence) over reciprocity-related traits (such as perceived warmth), this cultural dynamics maintains in the population both the less as well as the more polite variants in cycles of evolutionary change.

2 Responsibility exchange theory

For introducing responsibility exchange theory (RET), C&L invite us to imagine a common workplace scenario in which Olivia and Roger collaborate over a project, and Olivia contributes a key revision to the report that Roger subsequently hands in to the Supervisor, who then praises him for the report. Roger and Olivia then have a chance to communicate in a sequence of two steps: Roger may thank Olivia, thus transferring the credit for the success to her, but if he does not, Olivia still has the opportunity to claim the credit by bragging. This order is reversed if the Supervisor criticizes the report, resulting in the distribution of blame rather than credit: now in the first step, Olivia may apologize to Roger, and if she does not, in the second step Roger may choose to blame her for the failure. As a result of thanking, bragging, apologizing or blaming, the Supervisor now believes Olivia to be more, and Roger less responsible, relative to the Supervisor’s original belief about the distribution of the credit/blame.

C&L model both situations in game-theoretic terms in form of so-called extensive-form games.¹ Since the underlying game structure applies to any similar situation involving a transfer of responsibility for a positive or negative outcome, the initials O and R extend to the roles of Originator and Receiver, the former (mostly) responsible

¹ For exemplary game representations adapted from C&L see Appendix A.1 Fig. 5.

Table 1 A schematic presentation of the responsibility exchange theory. Adapted from Chaudhry and Loewenstein (2019)

	Originator (O)	Receiver (R)
<i>Credit situation: positive outcome</i>		
Communication	Bragging	Thanking
Transfer of responsibility	O takes credit from R	R gives credit to O
Competence and warmth	Gains competence, loses warmth	Loses competence, gains warmth
<i>Blame situation: negative outcome</i>		
Communication	Apologizing	Blaming
Transfer of responsibility	O takes blame from R	R gives blame to O
Competence and warmth	Loses competence, gains warmth	Gains competence, loses warmth

for the outcome, the latter receiving (most of) the credit. S extends to Spectator, which may be real or virtual (VS). C&L refer to research showing that people behave as if someone is watching them even when no one is, which allows C&L to explain how thanking, bragging, apologizing and blaming continue to have an effect even with no onlookers: although the actual distribution of responsibility is supposedly clear to the interacting dyad, the four communications still serve the function of informing the Virtual Spectator.

The crux of RET is in the two proposed valued social commodities, perceived competence and perceived warmth, which jointly form the social image of a person and underwrite the communicators' image-based utility functions. Thanking, bragging, apologizing and blaming "introduce image-based costs and benefits for both the communicator and the recipient of communication: Each of the four communications involves a tradeoff between appearing competent and appearing warm" (C&L). Giving away credit (thanking) or taking over blame (apologizing) makes the speaker appear less competent while increasing the perceived competence of their interactant. However, these two polite communications of thanking and apologizing, while decreasing the speaker's competence, act to increase her warmth. Conversely, claiming credit (bragging) or passing on blame (blaming) act to increase the speaker's perceived competence and decrease that of the interactant, which however happens at the cost of warmth, which the speaker loses through blaming or bragging (see Table 1).

Finally, the social value of competence versus warmth will not be uniform across all contexts, speakers and groups, which may all vary on a scale between warmth-favoring and competence-favoring. C&L formalize this as γ , essentially a weight value on warmth, so that competence is favored when $\gamma < 1$ and warmth is favored when $\gamma > 1$. Again, different situations may be conducive to different desirability of competence versus warmth, such as professional versus familial settings, or there may be stable individual preferences, so that, for example, the Originator and the Receiver may each have a different γ value. Most interestingly from our perspective, the relative values of the two aspects of social image may also vary between entire societies, a fact

that C&L use to explain the differences between collectivist and individualist cultures, such as the greater value of apologies in the former.

An additional point to bear in mind is that the labels of thanking, apologizing, bragging and blaming refer to functional rather than linguistic categories, whose definitions are tied to transferring credit or blame rather than specific linguistic formulations. This is clearly visible in the operationalizations in C&L's game and live chat experimental post-study, where the four categories are operationalized (perhaps a bit circularly) as *any communications* that function to acknowledge a good/bad performance of oneself/the other player (p. 329). This functional approach that abstracts away from implementation details is an important point to which we return to in Sect. 4.²

3 A strategy-based model

Let's revisit some central aspects of C&L's sequential model, based on epistemic game theory, and turn it into a cultural-evolutionary model by applying tools from evolutionary game theory on a normal-form game representation. Like C&L, we consider two agents whose communications affect their social image, so that in each situation an agent can gain or lose an amount of perceived competence $c \in \mathbb{R}$ in $(0, 1)$ and perceived warmth $w \in \mathbb{R}$ in $(0, 1)$. In this scenario, agents use politeness strategically: Being rational players, they choose to be or not to be (im)polite with the aim of maximizing their total image-based utility, which comprises competence and warmth.

3.1 A normal-form game representation of the RE scenario

Let us take a look at the credit situation with a positive outcome (see Table 1, top): The person mainly responsible for the outcome (originator, O) can either brag (B) or be quiet (Q), and the receiver R can either thank (T) or be quiet (Q), so the set of the originator's strategies is given as $S_O = \{B, Q\}$, and the set of the receiver's strategies is given as $S_R = \{T, Q\}$. The originator's utility function $U_O : S_O \times S_R \rightarrow \mathbb{R}$ defines the change in social image of the originator for strategy profiles consisting of an originator's and a receiver's choice. These changes are defined as follows: When the originator chooses B , she gains social image in terms of competence (c) and loses social image in terms of warmth (w), independently of the receiver's choice. Thus the originator's utility function U_O for playing B is given as $U_O(B, T) = U_O(B, Q) = c - w$. Furthermore, when the originator chooses Q , she receives c when the receiver thanks her and 0 else, thus $U_O(Q, T) = c$ and $U_O(Q, Q) = 0$. Similarly, the receiver's utility function $U_R : S_R \times S_O \rightarrow \mathbb{R}$ is defined as follows: When the receiver chooses T , he gains social image in w , but loses social image in c , thus his utility function U_R for playing T is given as $U_R(T, B) = U_R(T, Q) = w - c$. On the other hand, when the receiver chooses Q , he gets $-c$ when the originator brags and 0 else, thus $U_R(Q, B) = -c$ and $U_R(Q, Q) = 0$.

² Specific examples can be retrieved from C&L's online appendix, coding and sample chat transcripts.

Table 2 The responsibility exchange (RE) scenario as normal-form game representation, where the strategies are labeled with respect to the credit situation with the row player is originator and the column player is receiver

	T	Q
B	$c - w, w - c$	$c - w, -c$
Q	$c, w - c$	$0, 0$

The table can be adjusted to represent the negative outcome (blaming) scenario by replacing bragging (B) with blaming and thanking (T) with apologizing, and assuming the row player to be the receiver and the column player to be the originator

The resulting utility table with the originator as the row player and receiver as the column player is given in Table 2.^{3,4} Since the negative outcome scenario, in which the originator can blame or be quiet, and the receiver can apologize or be quiet, produces exactly the same table (when B stands for blaming, T stands for apologizing, and the roles for originator and receiver are interchanged), the following analysis of Table 2 covers both the credit and the blame situation. Therefore, we will refer to it more generally as the responsibility exchange (RE) scenario, although we will continue to refer to both actions as thanking (T) and bragging (B) for simplicity.

The RE scenario as given by the game of Table 2 adopts crucial aspects of the game(s) by C&L: it has the same strategy set for originator and receiver, and particular strategy combinations have a particular impact on her image in competence and warmth. However, there are also fundamental differences. First of all, the game of Table 2 is a so-called *normal-form game* representation, whereas C&L reconsider an *extensive-form game* representation (cf. Appendix A.1, Fig. 5). In a normal-form game, players are assumed to make their choices simultaneously, whereas in an extensive-form game, players can make their choices sequentially and thereby have information on the previous move of the other player. The normal-form game representation has the advantage of being more flexible⁵: in particular, it facilitates the application of

³ Note that in comparison with C&L there is a slight difference in the way these values are changed for particular situations. For details see Appendix A.2.

⁴ Note that the combination of bragging and thanking, B and T , has the same effect on the competence image as only bragging or only thanking alone: a transfer of magnitude c from row to column player. This is in the spirit of the scenario by C&L where playing B and/or T restores the original responsibility by revising the competence image by magnitude c . However, in other situations, instead of restoring an original value, bragging and thanking might independently affect the competence image, so that the result of the joint action would be a boosted transfer of a value greater than c . Such a boosted responsibility exchange (BRE) scenario can be defined as a generalization of the RE scenario. A definition of the BRE scenario and an evolutionary analysis is presented in Appendix A.3.

⁵ The limitation of the normal-form representation is ignoring the sequential order of the players' moves. However, note that C&L do not favor a particular order, but study both possible orders, where either the originator or the recipient makes the first choice. They show that both orders produce the same result under the assumption that both participants share the same preference order (as assumed in our model) and the extensive form game has at least a length of 3 (p. 323). This result shows that a particular order is generally not an essential aspect for the outcome of the scenario under discussion, and we think it pays off to drop it for the sake of more generality and a greater applicability of game-theoretic tools.

very important game-theoretic solution concepts, such as the Nash equilibrium (NE)⁶ or the evolutionarily stable strategy (ESS)⁷.

Note also that the game of Table 2 admits more strategy combinations than the original by C&L (cf. Appendix A.1, Fig. 5). Here, the joint option of bragging and thanking is possible, but excluded in C&L's model. Thus, we allow for all possible strategy combinations, even though some strategy profiles, such as the combination of bragging and thanking, seem to be odd. On the other hand, this combination is reasonable when we think of a scenario where originator and receiver each has a private face-to-face meeting with the boss, where one player does not know what the other would do. Further, the function of a good game-theoretical model is not to exclude unusual or non-rational strategy profiles a priori: rather, it is precisely the job of the model and its solution concept to determine and explain which strategy profiles are rational and which are not. For example, a rational agent would never brag when he knows that her opponent thanks, since $c > c - w$. Thus, although the game might contain non-intuitive strategy profiles, these are expected to be eliminated by general theories of rationality or—as it is important in the subsequent analysis—evolutionary dynamics.

Finally, Table 2 highlights two important aspects of the RE scenario: (i) competence is a zero sum relational good (whenever one player loses c , the other player gains c), whereas (ii) warmth is ultimately tied to the strategic choice (T gains w , B loses w). As we will discuss in Sect. 4.1, the first aspect does not necessarily hold for conversational scenarios beyond the RE scenario. However, in this section we will analyze this particular RE scenario type in terms of evolutionary game theory to study basic conditions for the stability aspects of politeness strategies in populations of interactive agents.

3.2 A population-based view

At this point, we are in a position to make a step from one-off interactions between individuals to applying this reasoning to a population of agents who alternate between the roles of originators and receivers. We assume that an agent's strategy is given by a strategy pair that defines how an agent would behave as an originator, and how an agent would behave as a receiver. For example, an agent that brags (plays B) when being in the originator role, and is quiet (plays Q) when being in the receiver role is defined by the strategy pair $\langle B, Q \rangle$.

Formally, an agent's strategy is defined by a strategy pair $\langle s_o, s_r \rangle$ of an originator strategy $s_o \in S_O$ and a receiver strategy $s_r \in S_R$. The strategy pair utility function $U_P : (S_O \times S_R) \times (S_O \times S_R) \rightarrow \mathbb{R}$ for an agent playing a strategy pair $\langle s_o, s_r \rangle$ against

⁶ A Nash equilibrium (NE) is a central concept in (epistemic) game theory. A strategy profile forms a NE when no player can improve her payoff by a unilateral change to another strategy. Moreover, a strict NE is a refinement of an NE. Here, any player's payoff strictly declines when unilaterally changing to another strategy. See Myerson (1991) for the formal definitions. For example, with respect to Table 2, it is easy to see that when $w > c$, then the strategy profile (Q, T) is a strict NE, since then $c > c - w$ and $w - c > 0$.

⁷ An evolutionarily stable strategy (ESS) is a central concept in evolutionary game theory. It will be informally introduced in the following subsection. For a formal definition see Maynard Smith and Price (1973).

Table 3 The symmetric RE scenario defined as utility table over strategy pairs of originator and receiver strategy and deduced from the (asymmetric) RE scenario of Table 2

	$\langle B, T \rangle$	$\langle B, Q \rangle$	$\langle Q, T \rangle$	$\langle Q, Q \rangle$
$\langle B, T \rangle$	0	0	0	0
$\langle B, Q \rangle$	$-w$	$-w$	$c - w$	$c - w$
$\langle Q, T \rangle$	w	$w - c$	w	$w - c$
$\langle Q, Q \rangle$	0	$-c$	c	0

an agent playing a strategy pair $\langle s'_o, s'_r \rangle$ is defined by the sum of originator utility U_O and receiver utility U_R as follows:

$$U_P(\langle s_o, s_r \rangle, \langle s'_o, s'_r \rangle) = U_O(s_o, s'_r) + U_R(s_r, s'_o)$$

The switch from defining agents by either originator strategy or receiver strategy to defining agents by strategy pairs of the two kinds enables us to symmetrize the game. Note that the RE scenario as presented in Table 2 is an asymmetric game, since the row player has different strategies than the column player. The game table over all strategy pairs defined by the strategy pair utility function U_P , however, is a symmetric one, where row and column player strategies are identical. Table 3 shows the resulting symmetrized game table of the RE scenario. Note that the table solely displays the utility values of the row player⁸, which is sufficient for symmetric games. The reason for symmetrizing the game table is to enable an intuitively more realistic analysis. For example, an asymmetric game table, such as Table 2, must be analyzed by a two-population model, where one population consists of originators, and the other of receivers. The symmetric table, such as Table 3, can be analyzed by a one-population model consisting of members that have two strategies: one for being in the originator position, and another one for being in the receiver position. Therefore, in the following we will use the Table 3 representation for the evolutionary analysis and address it as the symmetric RE scenario.

Note that the strategies are still labeled with respect to the positive outcome (credit) situation. To have more general strategy labels that also incorporate the blame situation, let us relabel the four different strategy pairs as follows: $\langle Q, T \rangle$ as a polite strategy P , $\langle B, Q \rangle$ as an impolite strategy IP , $\langle B, T \rangle$ with AC for ‘always communicative’ and $\langle Q, Q \rangle$ with AQ for ‘always quiet’. Adapting this to the blame situation, P stands for solely apologizing, IP for solely blaming, AC for both blaming and apologizing, AQ for doing neither. Table 4 shows the relabeled utility table of the symmetric RE scenario.

An evolutionary game theory (EGT) interpretation of the game assumes that we have a population of individuals, where each individual is associated with a particular

⁸ This is common practice for symmetric games, since the column player’s utilities are the same, when mirroring them at the top left to bottom right axis of the game table. For convenience, the complete game table with both players’ utilities is given in the Appendix A.4, Table 15.

Table 4 The symmetric RE scenario with generalized strategy labels

	AC	IP	P	AQ
AC	0	0	0	0
IP	$-w$	$-w$	$c - w$	$c - w$
P	w	$w - c$	w	$w - c$
AQ	0	$-c$	c	0

strategy.⁹ This situation can be represented by a population state. Let's for example assume that we have a population of 50% AC players and 50% IP players (P and AQ players are absent from this example, or in other words they each have a population share of 0%). This population state can be represented by a vector $(0.5, 0.5, 0, 0)$, where the first value stands for the proportion of AC players, the second for the proportion of IP players, and so on.

Given the utility table and a population state, we can compute the *fitness* of each strategy, defined as the expected utility of a strategy with respect to the given population state. Let's for example compute the fitness values for population state $(0.5, 0.5, 0, 0)$: As observable from Table 4, AC players score 0 against themselves and 0 against IP players. Thus, in the given population state AC score $0.5 \cdot 0 + 0.5 \cdot 0 = 0$ on average. Likewise, IP players score $(0.5 \cdot -w) + (0.5 \cdot -w) = -w$ on average. In other words, for the population state $(0.5, 0.5, 0, 0)$ where only the two strategies AC and IP exist, AC has a fitness of 0 and IP has a fitness of $-w$.¹⁰ Thus AC has a higher fitness than IP and would replace its competitor under evolutionary dynamics, ultimately leading to a population of only AC players, represented by the population state $(1, 0, 0, 0)$.

But now assume that only one P agent joins the population, so that it has a miniscule population share of ϵ . This would change the population state to $(1 - \epsilon, 0, \epsilon, 0)$. With the same calculation as before, we compute that AC has still a fitness of 0, but P has a fitness of w . Since $w > 0$, P would replace AC completely under evolutionary dynamics. In other words, a whole population of AC players can be invaded by an arbitrarily small number of P players.

This example demonstrates that AC is not an *evolutionarily stable strategy* (ESS), formally introduced by Maynard Smith and Price (1973). An ESS is a strategy that cannot be invaded by an arbitrarily small number of players with another strategy. In other words, it has an *invasion barrier* against all other strategies. Strategy AC is therefore not an ESS, since it does not have an invasion barrier against strategy P .

For the mathematical definition of ESS we refer the reader to the original paper by Maynard Smith and Price (1973). Here it is sufficient to say that if a strategy is a strict Nash equilibrium (cf. Myerson 1991) against itself, then it is an ESS. This can be easily spotted in symmetric tables, such as Table 4, by simply checking if the utility value of a strategy against itself is the unique maximum in the column. For example, strategy P scores w against itself. When we assume $w > c$, then we can see that w is the unique maximum in the column, since $w > 0$, $w > c - w$ and $w > c$. In other

⁹ Alternatively, it is possible to abstract away from individuals and interpret the population as a population of strategies.

¹⁰ It is not important if values are 0 or positive, but only the relation between values.

words, when $w > c$, then strategy P is a strict Nash equilibrium and an evolutionarily stable strategy: it cannot be invaded by an arbitrarily small number of copies of any other strategy.

3.3 An evolutionary analysis

An interim analysis shows that the game dynamics will differ depending on the relationship between the social value of warmth versus competence¹¹, which for simplicity we now assume to be a fixed value in a population at a given point in time. When $w > c$, thus warmth is more valued than competence, then P is the only ESS of the game table. In other words, in warmth-favoring populations it is expected that everyone thanks/apologizes and nobody brags/blames, since i) it is the only rational strategy due to being a strict NE, and ii) it is the only attractor state under evolutionary dynamics due to being the unique ESS.

For contexts favoring competence over warmth ($w < c$), the situation is more complicated. None of the four strategies is strictly dominated. Moreover, none of the four strategies is a Nash equilibrium against itself and therefore none is an ESS. In fact, evolutionary dynamics would produce a cyclic behavior of one strategy replacing the others in the following way: $AQ \rightarrow IP \rightarrow AC \rightarrow P \rightarrow AQ$. This means that in populations where competence has a higher value than warmth, we would see a ‘Cycle of Politeness’.

Figure 1 depicts the ‘Cycle of Politeness’ covering both the positive outcome and the negative outcome scenario¹²: In an AQ population, being impolite (here: bragging or blaming) is beneficial by value $c - w$, since one claims the more valuable currency of perceived competence at the cost of losing the less valuable currency of perceived warmth. This leads to an impolite IP population. In an IP population, being polite (thanking, apologizing) is beneficial by a value w , as it does not harm perceived competence—since everyone is claiming competence anyway—and at the same time increases perceived warmth. This leads to an AC population. Now, in an AC population it is beneficial to give up being impolite by a value w , as it does not harm perceived competence—because the other person attributes it to the originator—and at the same time increases perceived warmth. This leads to a polite P population. Finally, in a P population it is beneficial to give up being polite by a value $c - w$, since one gains c by not attributing it to the other person, in exchange for losing w . This leads to an AQ population.

It should be emphasized that the cycle shown above represents only an idealized extreme case of the entire temporal dynamics. It assumes that at one point in time all agents are in a AQ state and get all replaced by IP agents, which in turn get all replaced by AC agents afterward, and so on. But the evolutionary dynamics of the game can produce multiple orbits of cycles among mixed population states. For

¹¹ As already mentioned, C&L used a weight factor γ to control the relationship between perceived warmth and perceived competence, because they define both image types as functions. Since we define them as update values, we can show them as directly related.

¹² This figure also highlights the fact that the cycle only works if $c > w$. For $w > c$ the transitions would all run to a P population, since the arrows from P to AQ and from AQ to IP would be reversed.

Fig. 1 Politeness cycle among the four stages AQ , IP , AC and P . Transitions are labels with the gain from switching from one to the other strategy

Table 5 Utility table of the RPS game

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

example, if there is a population of 50% AQ agents, 40% IP agents and 10% AC agents, then the AQ agents get replaced more and more by IP agents, but at the same time the already existent IP agents get replaced by AC agents.

To better illustrate the cyclic behavior of the game table (Table 4) under evolutionary dynamics, we refer to a well-known game that reproduces similar dynamics for a 3×3 strategy space: the Rock-Paper-Scissors game (RPS, Table 5). RPS is a zero-sum game where one player wins (1 point) when the other loses (-1 point), or both get nothing for a tie when they choose the same (0 points). Here rock (R) beats scissors (S), scissors beats paper (P), and paper beats rock. This is parallel to our Politeness game, where IP beats AQ , AC beats IP , P beats AC and AQ beats P .¹³

¹³ For completeness it should also be mentioned that P beats IP when $c < 2w$, and IP beats P when $c > 2w$.

Fig. 2 The trajectories of the RPS game under replicator dynamics projected on a simplex

Figure 2 represents the temporal dynamics of the RPS game under the well-known *replicator dynamics*¹⁴ projected on a simplex. This simplex represents the vector field of population states, and the arrows in the simplex are sample gradients that represent the directions of change under the replicator dynamics, whereby the length of an arrow shows the velocity of change. Therefore, by following the arrows one can derive the trajectories of change. For example, let's start with the top point of the simplex that represents a population where everybody plays *R*. This population can get invaded and completely replaced by *P* players, reaching the bottom left point where everybody plays *P*. Now, this population can be easily invaded by *S* players until the bottom right point is reached where everybody plays *S*, and so on. But when we reconsider population states of the interior area of the simplex, the dynamics still follows the cyclic behavior, but now for mixed population states consisting of all three strategies.

A very special case for the dynamics behavior of the RPS game is given when the population plays each strategy for the same share: $\frac{1}{3}$ play *R*, $\frac{1}{3}$ play *P* and $\frac{1}{3}$ play *S*. This population state is a rest point (marked by the dot in the center of the simplex in Fig. 2) under evolutionary dynamics without mutation. This means that once a population reaches this state, it stays in this state. But this rest point is not evolutionarily stable: any slight mutation would make the population state spin in a counter-clockwise fashion. Moreover, under replicator dynamics this spin would lead the cycle more and more to the border of the simplex: approximately 100% of the population would stay a particular time in a state of using only one strategy, until it is almost completely conquered by a strategy that beats it, and stay there for an even longer time. These time intervals grow progressively longer.

¹⁴ The replicator dynamics, introduced by Taylor and Jonker (1978), is a standard dynamics in EGT. Simply put, it follows the principle that a strategy's population share increases (decreases) if and only if its fitness is better (worse) than the average fitness across all strategies with respect to the given population state.

Fig. 3 Changes of population shares over time under replicator dynamics for the game of Table 4 with $c = 2w$. Starting point: $f(AC) = f(IP) = 0.24$ and $f(AQ) = f(P) = 0.26$

Exactly the same applies to the Cycle of Politeness, but with four strategies. Let us reconsider the game table of Table 4 with $c = 2w$. Furthermore, let's define $f(s)$ as the population share of a strategy s . Then for the given game table a population with population shares of $f(AQ) = \frac{1}{4}$, $f(IP) = \frac{1}{4}$, $f(AC) = \frac{1}{4}$ and $f(P) = \frac{1}{4}$ is a rest point where nothing changes under the replicator dynamics. But for a slight shift away from this population state, we would see a cyclic behavior where each strategy gets longer and longer phases to dominate the whole population until it gets replaced by the next one of the cycle in Fig. 1.

Figure 3 shows the change of the population share for all four strategies under the replicator dynamics for the game table of Table 4 with $c = 2w$, by starting with an initial population state very close to the rest point, namely of $f(AC) = f(IP) = 0.24$ and $f(AQ) = f(P) = 0.26$. As we can observe, the strategies replace each other in the order illustrated in Fig. 1, and for each majority strategy, its population share continues to grow until it approximates 100%, and then each majority strategy's reign until replacement becomes increasingly longer over time.

Importantly, the fact that the population state successively approaches the border of the simplex is an artefact of the particular dynamics (the replicator dynamics). Other evolutionary dynamics might produce a movement to the central rest point, still others might stick in the same orbit. However, what is invariant across all evolutionary dynamics is that the change in population state follows the same direction, namely the one depicted in Fig. 1.

Another important result of the analysis is that for the example in Fig. 3, where $c = 2w$, the average population share of each strategy over time approaches $\frac{1}{4}$ each when time goes to infinity.¹⁵ In other words, when $c = 2w$, then we have a Cycle of Politeness with no temporally dominant strategy, since each strategy is expected to be played with the same frequency over an infinitely long interval. The situation changes when $c \neq 2w$. More precisely, it can be shown that for $c < 2w$, the strategy

¹⁵ See Appendix A.5 for a representation of average population shares over time.

Fig. 4 Result of the evolutionary analysis for different c to w ratios

AQ is temporally dominant at the expense of strategy AC , whereas when $c > 2w$, the strategy AC is temporally dominant at the expense of strategy AQ .¹⁶

Figure 4 summarizes the overall result of our analysis. First of all, we know from the last section that P is the only ESS and strict Nash equilibrium when $c < w$. In other words: for situations that favor warmth over competence, polite behavior is the only culturally expected as well as rationally justified strategy. On the other hand, when $c > w$, then evolutionary dynamics produces a ‘Cycle of Politeness’ with the order $AQ \rightarrow IP \rightarrow AC \rightarrow P \rightarrow AQ \rightarrow \dots$ When $c = 2w$, then there is no dominant strategy and the cycle gives every strategy the same amount of population share over an infinite time interval. But when $c > 2w$, then the temporally dominant strategy of the cycle is AC , whereas when $w < c < 2w$, then the temporally dominant strategy of the cycle is AQ . To sum up,

1. When social image in warmth is more valued than social image in competence ($w > c$), then the “polite” strategy P is the only strict NE and also the only evolutionarily stable strategy.
2. When social image in warmth is less valued than social image in competence ($w < c$), then none of the four strategies is a NE or evolutionarily stable, but the situation represents a RPS-like zero-sum scenario, where evolutionary dynamics produce a ‘Cycle of Politeness’.

3.4 An interpretation of the results

As the results of the evolutionary analysis showed, we have to distinguish among two different situations of the RE scenario. In one situation, when $w > c$, the Politeness

¹⁶ See Appendix A.6, Fig. 7 for the representation of the change of population share with $c = \frac{3}{2}w$, and Fig. 8 for the representation of the change of population share with $c = 3w$, both under replicator dynamics.

strategy P (apologize as a bad originator, or thank the good originator) is evolutionary stable. In such situations everyone is expected to be polite, and being polite is the best strategy in such a polite environment. In the other situation—when $w < c$ —none of the strategies is evolutionary stable. People may be polite or not, impolite or not, and this may change from one situation to another. In such a scenario, the best strategy depends on the expectation of how the others behave. When people can be expected to be mostly polite, the best strategy would be to be mostly quiet. When people can be expected to be mostly quiet, the best strategy would be to be mostly impolite, and so on (cf. Fig. 1).¹⁷ In other words, here every strategy is rationally as well as evolutionarily plausible, and we would see and expect polite as well as “neutral” and impolite behavior.

The result that P is not evolutionarily stable in all situations is fundamental to the strategic character of the use of politeness. If we only had situations with $w > c$, politeness would lose its effect: since P is an ESS in any case, it would be universally used in the population as the only strategy, leading to a potentially infinite growth in everyone’s social image (here: warmth). But like material wealth, social image is relative in that it only works in comparison to the other members of the population: if everyone is rich, nobody is, and similarly, if everyone increases her social standing, nobody does. If we reconsider a more realistic model, where an agent’s increase in social image happens in relation to her total social image, then when everyone plays P , everyone increases her social image by the same percentage, and as a result no one’s social standing goes up relative to all the others. Using P uniformly does not have any long-term effect, since every increase is mitigated by immediate inflation.

When $w < c$, we have a cyclic dynamics which changes the overall behavior of the population over time. It implies that the polite strategies of thanking and apologizing, or their less polite counterparts of bragging and blaming, pay off in some population states, but do not pay off in others. It is essential that the valuation $w < c$ creates a conflict of interests between the two parties, since one has interest in revealing the actual responsibility behind the positive/negative outcome, while the other does not. This conflict of interests is what keeps the system dynamic and makes being polite both costly (by value c) and beneficial (by value w), depending on the population state. Polite behavior has an impact on social image that changes over time, creating a pressure on agents to adapt to this dynamics: being polite, impolite, both, or none, can each provide the best payoff at different points in time, which prevents the extinction of any of the variants. This makes the more polite strategies an option that a person CAN but does not HAVE TO choose, and having reasonable alternatives is exactly what gives politeness its *strategic* character.

4 Discussion: politeness and reputation

The analysis we presented above remains considerably limited by the original setting of the Responsibility Exchange Theory. C&L confine RET to a narrow set of situations

¹⁷ This is similar to a Rock-Paper-Scissors scenario where the opponent plays Scissors disproportionately often. Under such circumstances the best strategy would be to play Rock.

involving a dyadic transfer of responsibility for a non-zero material outcome. Having such a specific focus is what lets RET achieve strong evidential support: it allows C&L to establish a robust empirical basis for their assumptions, then complement their theoretical proposal with a formal game-theoretic model, and finally verify its predictions through an experiment. We note however that this is achieved at the cost of the theory's scope, and that the same approach and a set of tools, in particular the constructs of competence, warmth and Virtual Spectator, appear to have explanatory and possibly also predictive power that extend beyond the narrow confines of RET. In sum, C&L's approach holds considerable untapped potential, and we devote the remaining sections to outlining a sequence of steps in which the theory can be enhanced both in breadth and in depth.

In the first step, we demonstrate that the essence of RET extends beyond cases involving only an exchange of responsibility, which we do by upgrading C&L's original model to cover the signaling of another unobservable quality. Having established this, we venture more general suggestions on further extensions: in breadth, towards a more encompassing account of strategic conversational politeness, and in depth, by delineating how the currencies of (perceived) warmth and competence can be grounded in an ultimate-level account of reputation. Then we show ways of how the underlying mechanisms of our approach can be connected to other conversational phenomena, particularly (polite) requests and honorifics. Finally, we point to diachronic investigations of politeness and its change, and we conjecture possible congruence with the diachronic predictions of our model.

4.1 Going beyond a transfer of responsibility

In our first step, we show that the general ideas behind RET, and the main mechanics of its conceptual works, can be applied to scenarios beyond RET. Our point of departure is the observation that the use of thanking and apologizing specifically around a transfer of responsibility does not seem to be a category intuitively distinct from other uses, and it is not clear that it should work in some way qualitatively different from thanking and apologizing at large. Indeed, C&L's own examples (pp. 316–317) are a mixture of cases that RET can and cannot explain, without the two classes appearing different in other respects. We will elaborate on this point with an additional scenario described by the following two situations.

- *Situation 1:* John goes to a restaurant for lunch with friends. The ambience and the food are very good, and the waiter does an excellent job. John is really happy. When he is about to leave, he might or might not thank the waiter for the great service. On the other hand, the waiter might or might not brag about the own performance. What behavior would you expect to be more likely?
- *Situation 2:* John goes to the same restaurant for dinner with his partner. The restaurant is literally packed, the ambience is noisy and the food is almost cold when it arrives. The waiter forgets two times to bring the wine, and after the meal John has to ask for the bill twice. John is really unhappy. When he is about to leave, he might or might not complain to the waiter for all the inconveniences. On

Table 6 The RE scenario vs. the OS scenario: in the RE scenario, the quality of the outcome is common knowledge, and the true responsibilities must be signaled via politeness strategies

	RE scenario	OS scenario
True responsibilities:	Must be signaled	Common knowledge
Quality of outcome:	Common knowledge	Must be signaled

On the other hand, in the OS scenario, the true responsibilities are common knowledge, and the quality of the outcome must be signaled via politeness strategies

the other hand, the waiter might or might not apologize for the bad experience. Again, what behavior would you expect to be more likely?

The restaurant scenario has particular things in common with the RE scenario. There is a good outcome situation (situation 1) where the originator (representative of the restaurant) has the option to brag and the receiver (customer) has the option to thank; and there is a bad outcome scenario where the originator has the option to apologize and the receiver has the option to complain (which equals blaming in the RE scenario). However, there is also a fundamental difference. In the RE scenario, the outcome is common knowledge, but the responsibilities for the outcome are not common knowledge, and the four im/polite communications are used to signal the actual attribution of responsibilities. In our restaurant example, it is exactly the other way around: the responsibilities for the service are common knowledge (nobody would make the customer responsible for a restaurant's bad service), but the outcome is not, and the im/polite communications are used to signal the quality of the outcome. Therefore, we will refer to the restaurant example as the outcome signaling (OS) scenario. Table 6 shows a juxtaposition of the RE and OS scenarios.

The next step is to transfer the OS scenario into a normal-form game table in the same way as we did it with the RE scenario. In the good service situation, a bragging waiter will lose w , whereas a thanking customer will gain w . Moreover, only the waiter's (as representative of the restaurant) competence image is at stake. The waiter will gain c whenever the good service will become public knowledge, thus for any strategy pair, except when both parties are quiet. The resulting utility table is represented in Table 7 (left).

In the bad service situation, an apologizing waiter will gain w , whereas a complaining customer will lose w . Again, only the waiter's competence image is at stake here. Now the waiter will lose c whenever the bad service becomes public knowledge, thus for any strategy pair, except when both parties are quiet. The resulting utility table is represented in Table 7 (right).

On comparing the utility tables of the OS scenario and the RE scenario, we notice that in either case the effect on the warmth image is identical. Polite communications (thanking, apologizing) earn w and impolite communications (bragging, blaming, complaining) cost w . Note that we assume this property to be invariant for any further extension of the scenario space. What actually differs from scenario to scenario is the impact on the competence image. For example, in the RE scenario, c is transferred from one party to the other, whereas in the OS scenario c can only be attributed to one party (originator), and its sign (plus or minus) depends on the quality of the outcome.

Table 7 The outcome signaling (OS) scenarios as normal-form game representations

	T	Q		C	Q
B	$c - w, w$	$c - w, 0$	A	$w - c, -w$	$w - c, 0$
Q	c, w	$0, 0$	Q	$-c, -w$	$0, 0$

Situation 1: good service Situation 2: bad service

The table on the left represents situation 1 (good service), the table on the right-situation 2 (bad service). The row player is the originator (the waiter as a representative of the restaurant), the column player is the receiver (the customer) in both tables. The strategies are bragging (B), thanking (T), and being quiet (Q) in the first situation, and apologizing (A), complaining (C), and being quiet (Q) in the second situation

A detailed analysis of both tables with respect to Nash equilibria and evolutionarily stable strategies is given in Appendix A.7 and follows the same approach as we applied in the last section. The results are as follows. For the good service situation, there is exactly one strategy that is NE and ESS, independent of $w > c$ or $c > w$, and that is the strategy pair $\langle Q, T \rangle$, represented as the polite strategy P . In other words, when the service was good, the only rational and culturally expected strategy would be to thank the waiter, whereas we would never expect to see the waiter bragging. In the bad service scenario the result is more complex. Here we have to distinguish between $w > c$ and $c > w$. More concretely, when $w > c$ then the strategy pair $\langle A, Q \rangle$ is the only NE and ESS. Thus, when warmth is favored over competence, the only rational and culturally expected strategy would be that the waiter apologizes for the bad service. However, when $c > w$ then $\langle Q, Q \rangle$ is the only NE and ESS. Thus, when competence is favored over warmth, the only rational and culturally expected strategy would be for both parties to remain silent.

These results are largely in accord with our intuitions. Ask yourself: When was the last time you saw a waiter bragging about his service? Probably never. We surely see very frequently customers thanking for good service and waiters apologizing for bad service. This is all predicted by the model. But sometimes we also see customers complaining about bad service. According to our model, this strategy is not a rational or culturally expected choice with respect to the assumption of utility-maximizing agents. How, then, can we explain the existence of complaining customers?

The most likely explanation for such apparently non-rational behavior can be found in the principle of *strong negative reciprocity*, empirically attested as a property of the human evolved psychology¹⁸ and based on *fairness and punishment*. Let's transfer this principle to the bad service situation of the OS scenario. Note that when $c > w$ the equilibrium strategy pair is $\langle Q, Q \rangle$. However, when the waiter does not apologize, the customer might feel unfairly treated. The customer doesn't want to let the waiter get away with delivering a bad service and at the same time avoiding the costs of a bad reputation in competence. Thus, the customer is willing to pay a price by losing w to make the waiter to pay a higher price by losing c . Thus, strong negative reciprocity based on fairness and punishment can easily explain complaining customers. However,

¹⁸ See e.g. (Gintis 2000; Fehr et al. 2002). A famous game that demonstrates this principle for explaining seemingly non-rational behavior is the so-called *ultimatum game* (cf. Oosterbeek et al. 2004).

Table 8 The result of the game-theoretic analysis over eight different contexts, defined through three dimensions (i) RE or OS scenario, (ii) good outcome or bad outcome, and (iii) $w > c$ or $c > w$

	Good outcome		Bad outcome	
	$w > c$	$c > w$	$w > c$	$c > w$
RE scenario	P	cycle	P	cycle
OS scenario	P	P	P	AQ

While in more than half of all contexts (5 of 8) the polite strategy P is the only NE and ESS, the impolite strategy IP is never an NE or ESS

with the aim to keep our model simple and general, we tolerate that complaining is not part of our set of rational and evolutionarily stable solutions.

It is time to wrap up. All scenarios and situations discussed so far entail eight contexts, which differ with respect to three dimensions: (i) signaling the responsibility or the outcome (RE scenario vs. OS scenario), (ii) good outcome or bad outcome, and (iii) $w > c$ or $c > w$. Table 8 shows for all eight contexts which strategy is an NE and ESS, or if the result is a cycle, respectively. As the result shows, the polite strategy P is NE and ESS in more than half of all contexts, but not in every context. This result squares with the intuition that we find polite communications frequently and in many situations, but we do not expect people to be always polite. Moreover, communicating impolitely is never an evolutionarily stable strategy across all contexts.

4.2 Strategic conversational politeness

Can the explanatory apparatus of RET be further extended, not only onto additional contexts but also onto a larger class of instances of strategic use of polite language in conversation? This is a difficult task, but we devote the remaining sections to sketching out ideas on how the present approach can be developed towards a broader, functional and naturalistic account of SCP. In previous sections, we have already used the general labels “polite” versus “impolite” strategy following C&L’s observation (p. 314):

...communications that bolster another person’s perceived competence at the expense of one’s own perceived competence (i.e., thanking for a positive outcome or apologizing for a negative one) are seen as generous and polite. The opposite is true of blaming and bragging, which bolster one’s own competence at the expense of the other, and thus, are considered rude.

Of course, thanking and apologizing cannot be assumed as polite *by definition*: large bodies of Linguistic Politeness research (in particular “second wave”: Blum-Kulka 1992; Eelen 2001; Watts 2003; Kádár and Haugh 2013) have conclusively established that there is no strict correspondence between linguistic forms of illocutionary acts and politeness. However, thanking and apologizing as being polite *by default* is sufficient for statistical, *ceteris paribus* generalizations concerning SCP—the interface layer of conversation that is intermediate between the linguistic implementation and social payoffs. Again, *strategic use of conversational politeness* (SCP) is a notion that we introduced in Sect. 1, with politeness understood in intuitive terms, and strategy in economic terms, amenable to a game-theoretic and optimality-theoretic treatment.

Importantly, SCP implies a functional individuation of categories, thus allowing us to bracket the complexities of linguistic realization, which can be very considerable. Furthermore, face-to-face conversation, “the core ecological niche for language” (Torreira et al. 2015), is multimodal, not just in the sense of sensory modalities, but more importantly in the sense of the semiotic modalities or semiotic systems involved. This means that the signals of SCP need not be exclusively verbal, as long as they achieve their strategic function in conversation. A good example of such a functional approach is the recent study by Floyd et al. (2018, p. 5) on the universality of expressions of gratitude, who operationalize this category as “...any positive conveyance of appreciation or satisfaction by the requester immediately after receiving a response indicating the fulfilment of the request”, and further confirm the inclusion of nonverbal signals: “we included conventional phrases like ‘thank you’ as well as other forms with a comparable effect like ‘good job’ or ‘sweet’; we also included explicit acknowledgements done non-verbally, for example by nodding one’s head or making a hand gesture.”

At the same time, it is easy to see why prototypical signals of SCP are linguistic, and also why polite language has prototypical realizations in the form of conventionalized politeness markers. The key points that our construal of SCP inherits from RET are that the function of polite communications is trading off social commodities (specifically, warmth versus competence in RET), and that it is implemented via informing the Virtual Spectator of unobservable qualities (specifically, the allocation of responsibility in RET). Fulfilling this function hinges on successfully updating the knowledge state of the VS, something that is most effectively achieved through the use of conventional expressions.

4.3 From social image to reputation: towards an ultimate-level account

Espousing an ultimate-level perspective instantly reveals a rather sizeable elephant in the room: *Why do people value (perceived) warmth and competence?* Despite references to evolutionary game theory and signaling theory, C&L’s explanations remain proximate-level in that respect. It is an intuitively clear and empirically confirmed fact that (perceived) competence and warmth constitute valued social currencies, and C&L do not further explore the ultimate-level, i.e. fitness-based, reasons behind their purchasing power. We think that such reasons, i.e. in what ways competence- and warmth-seeking agents are successful in human societies, are likely to hold answers to a more comprehensive understanding of the value of SCP. Below, we speculate about the ways of extending the proximate-level notion of *social image* we have used so far to a more ultimate-level and naturalistic notion of *reputation*, by linking:

1. competence to (a capacity for acquiring and) *holding resources*,
2. warmth to (a propensity for) *sharing resources*.

An important common denominator is that both acquiring and sharing resources fundamentally determine one’s desirability as a cooperative partner, so such a grounding of reputational payoffs invites making contact with general theories of cooperation, such as “the leading eight” Ohtsuki and Iwasa (2006) or biological markets theory Noë and Hammerstein (1995).

From a more ultimate-level perspective, we take perceived competence to be extensible to a more general notion of status, i.e. an individual's vertical position in a hierarchy, which highly correlates with its ability to *acquire and hold resources*. The literature on management, particularly the psychology of management, demonstrates that perceived high status strongly correlates with perceived competence, e.g. there is empirical evidence that we perceive higher status individuals as more competent and lower status individuals as less competent; or even that high status individuals perceive themselves as more competent than they actually are (the theory of overcompetence, e.g. Belmi and Pfeffer 2018). This puts a pressure on high status individuals to show their competence and hide their incompetence, or even mask their incompetence by presenting it as competence. Interestingly, high status individuals or groups are also perceived as less warm (Brambilla et al. 2010).

A more generalized notion of status (rather than just competence) is also close to how utility-based accounts of politeness use the notion *face*. Clark (2012) makes this link expressly in hypothesizing that “face developed from dominance hierarchies in primate groups” (p. 278), and a similar construal is implied in several accounts of polite requests, which assume that a polite request effects a transfer of face from the requester, who loses some of it, to the requestee, who gains some (Asher and Quinley 2012; Quinley 2012; Quinley and Ahern 2012; Mühlenbernd et al. 2019).

To Pinker (2007) it is negative face more specifically that is connected with social standing, as it reflects ‘status’, ‘power’, ‘authority ranking’, and ‘dominance’ (p. 380, 405). Danescu-Niculescu-Mizil et al. (2013) document an inverse correlation between politeness and several different types of status in large corpora of online texts, in that higher status individuals tend to be less polite. From a broader naturalistic perspective, status is fundamentally constitutive of fitness in social animals, which suggests that this mechanism might have analogs in nonhuman primates (e.g. pant-grunts in chimpanzees, Kutsukake and Hosaka 2015; Sakamaki and Hayaki 2015).

Alternatively, or on different occasions, perceived competence may be a proxy for prestige-based status more specifically (as opposed to dominance-based, see Henrich and Gil-White 2001), but thanking and other SCP signals would have a double function: first, marking the interaction as one taking place between social equals (Fiske 1992, 2004; Pinker 2007), and then, serving to transfer perceived competence within that context.

Warmth is a more intriguing trait that lacks an obvious ultimate grounding, but from C&L's construal of it as likability, generosity, prosociality and trustworthiness, we may speculatively extend it to a person's generalized propensity to share resources. The reasons why a reputation for sharing has consequences to fitness in humans, possibly uniquely in the animal kingdom, is the foundational importance of cooperation and reciprocity to human societies (cf. Tomasello 2008), which have a variety of mechanisms for keeping track of and excluding non-cooperators (cf. Nowak 2006). On this construal, thanking and apologizing may be an ancillary mechanism for monitoring direct and indirect reciprocity, and rather than serving to pay up a debt could serve to reaffirm its existence. This interpretation is directly at odds with C&L's proposal, but is backed up with persistent intuitions from a broad range of literature, that by thanking or apologizing one accepts a debt (Brown and Levinson 1978) or by requesting politely, one “incur[s] a social debt (with respect) to one's conversational partner”

Table 9 Two forms of reputation: *c*-based and *w*-based reputation

	<i>c</i> -based reputation	<i>w</i> -based reputation
Definition	Capacity for acquiring and holding resources	Willingness/propensity for sharing resources
Manifestations	Competence, confidence, dominance, status, power	Warmth, helpfulness, generosity, reciprocity, trustworthiness

(van Rooy 2003; Quinley 2012; Quinley and Ahern 2012). However, whereas C&L's account shows warmth-signaling to be costly and resistant to inflation, this latter construal would need to posit additional reputation-based monitoring mechanisms for keeping warmth signals stable against being exploited as cheap talk.¹⁹

To sum up, Table 9 presents a juxtaposition of *c*-based reputation and *w*-based reputation with respect to definitions and manifestations.

4.4 Extensions to other forms of politeness

As our final thought, we outline directions for extending the *w*- vs. *c*-reputation based accounts of LP phenomena beyond communications surrounding responsibility transfer, which we illustrate with polite requests and honorifics.

As we have mentioned above, requesting politely is intuitively perceived as involving social costs, an idea that has been implemented in the few existing game-theoretic models of polite requests (see Quinley 2012; Asher and Quinley 2012; Ahern and Quinley 2014; van Rooy 2003). Such models refer to the social currency *face*, and construe linguistic politeness devices as triggering the transfer of *c*-based reputation from the Requester to the Requestee. E.g. Ahern and Quinley (2014, pp. 6–7) write: “polite requests entail a loss of face on the part of the requester; so to speak, the requester makes a face ‘investment’ in the requestee [...] If X asks for help, using a polite request, Y should experience some boost in face...”. However, Ahern and Quinley (2014, p. 13) also observe how the use of politeness devices in requests can trigger a transfer of *w*-based reputation aspects to the Requester: “by following the norm of politeness, one indicates that one will also follow the norm of reciprocity. Being aware of and using the politeness forms of a particular group suggests a set of shared obligations...”.

¹⁹ An excellent candidate here is the notion of (perceived) *reliability*. See e.g. Briñol and Petty (2009) for the role of reliability/credibility and confidence/competence in persuasion strategies, or, more specifically, McCready (2015) for a game-theoretic study of (one's reputation for) reliability in pragmatics, in particular her treatment of hedges and evidentials (also see Giardini et al. 2019 for relevant empirical evidence). Here a person's reputation for reliability is connected to the level of truthfulness of her assertions. With respect to the discussed scenarios, reliability would not be strongly attached to either competence or to warmth, but would form an interesting third dimension. Let's for example take the RE scenario: An agent who wants to maximize warmth always plays *P* (thanks, apologizes), whereas an agent who wants to maximize competence always plays *IP* (brags, blames). However, an agent who wants to maximize reliability always plays *AC* (thanks, apologizes, brags, blames), since any form of not being quiet helps to reveal the true state of affairs on the RE scenario. On the other hand, an agent who always plays *AQ* might be considered as unreliable, since she happens to conceal the true states of affairs.

Similarly, van Rooy (2003, p. 55) refers to two types of social currency, which map to *c*-based and *w*-based reputation, respectively: “Thus, it are social costs that are at issue here. But what could these costs be? ... 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.” This analysis gains some general support from ethological observations as well as empirical data on actual language use, in particular in terms of *c*-based reputation transfers. Zahavi and Zahavi have proposed this to be a general property of human and non human communication [in the latter, utterance length is used as a proxy for politeness level (cf. Leech 1983, pp. 107–110; Östman 1989)]. They state that “the act of petitioning is costly to the requester. It also decreases the standing or prestige of the requester, and increases the standing of the giver, in the eyes of witnesses.” (Zahavi and Zahavi 1997, p. 75)

Another area of language use that documents the interface between status and politeness is the use of honorifics. McCready (2019, p. 2–3) gives the following as a standard definition of honorifics: “they are those expressions which perform the linguistic marking of ‘honorification: relationships involving social status, respect or deference between communicative interactants’ Agha (1994)”. She mentions a number of forms of honorification, such as honorific particles, terms of address, pronoun or register types. Without delving into the intricacies of the subject, it should be noted that irrespective of their specific characteristics, they can all be used as SCP, whereby the use or non-use of honorifics may assert a particular status relation between interactants.

For illustrative purposes, let’s look at terms of address, the treatment of which has a long history in sociolinguistics and LP studies (e.g. Brown and Gilman 1960; Brown and Ford 1961; Lakoff 1973, 1990; Brown 2017), including the research on how they reflect social configurations and norms (Lee-Wong 1999; Locher 2012). In non-reciprocal contexts—for example, the student addressing the teacher with an academic title—honorifics serve to underline the status difference between the speaker and the hearer. The teacher may accept such a configuration of status, in this way asserting his competence, or challenge it by suggesting the dropping of her title and in this way increase her perceived warmth. This basic dynamics of honorification applies to other contexts, such as the reciprocal situation, in which two professors either use or don’t use their academic titles (whereby they respectively underscore their competence or warmth), and to all other forms of honorification; for example, competence can be advertised by the use of higher register and warmth by the use of lower register.

Marcjanik (2015) provides an interesting example of recent changes in the Polish etiquette of using addressative terms. The traditional Polish etiquette demanded that in interaction between equals both parties should obligatorily use the honorifics *Pan* (‘Sir/Mr.’) or *Pani* (‘Madame/Mrs.’). In interactions between unequals the lower status party was expected to use an honorific acknowledging the higher status of the fellow interactant (e.g. Mr. Dean, Mrs. Professor, Mrs. Director), and the higher status party, the regular term of address *Pan/Pani*. The highly codified and normative character of the traditional Polish politeness required the use of adversative terms even in decidedly aggressive contexts (e.g. *Spierdalaj*, *Pan!* – ‘Fuck off, Sir!’). The social and economic changes instigated by the fall of communism promoted a move towards more egalitarian and less formulaic forms of politeness, which for example manifest themselves in a growing tendency to drop the hitherto obligatory *Pan/Pani* in contacts between

strangers or to drop honorifics when addressing a higher status party, particularly in professional context, and replace them with first-name vocatives Marcjanik (2015). Another consequence of this modern egalitarian attitude is bolstering the strategic use of linguistic politeness, which was constrained by the normative nature of the traditional Polish etiquette. For instance, in academic correspondence a higher status party can use a competence-oriented strategy and highlight a status difference by signing an email with a full array of academic titles on top of her first and last name (Prof. dr hab. Małgorzata Marcjanik), which was the standard course of action in the old days; or opt for a warmth-oriented strategy and sign it only with her first name (Gosia, the diminutive form of Małgorzata).

The examples given in this section illustrate the links between linguistic politeness and status, which can be interpreted in terms of competence and warmth. Of course, as such they do not directly support one main finding of our analysis, i.e. the cyclic nature of the evolution of politeness. Tracing down the cyclicities predicted by our model in actual populations is difficult because of several model-internal factors (e.g. initial frequencies of each strategy, individual variation in the utilities of w and c) and a very large number of model-external factors that affect social and linguistic change, such as economic stratification or language contact. Controlling for this wealth of factors may appear impossible in actual language use, which appears to be several orders of magnitude too noisy for the signal (i.e. the cycles) to be detectable. However, these obstacles should not prevent us from looking for relevant linguistic evidence in studies on historical changes in politeness systems.

Although the available studies on this topic are rare and usually theory-laden—e.g. see the studies into the evolution of English politeness by Watts (2003), Jucker (2012) or Leech (2014)—it still seems possible to distill from them some general trajectories of change and confront them with the predictions of our model. For instance, (Leech 2014, p. 286) argues that the Old English period favored a discernment-oriented forms of politeness, which includes the denigration of opponents and self-praise. The latter is illustrated by the passage from *Beowulf*, where the main protagonist brags about his prowess in battle:

These men knew well the weight of my hands. Had they not seen me come home
from fights Where I had bound five Giants—their blood was upon me. [lines
418–421]

In terms of the politeness stages of our model, this form of politeness would correspond closely to an IP phase, where bragging/blaming is received more favorably than thanking/apologizing.

The Middle English politeness, derived from the ideals of courtly speech, was according to (Leech 2014, pp. 286–288) characterized by deference, which could assume very elaborate forms, as illustrated by William Caxton's editorial prologue, where he addresses his patroness:

Unto the right noble, puissant and excellent princess, my redoubted lady, my Lady Margaret, Duchess of Somerset, mother unto our natural and sovereign lord and most Christian king, Henry the Seventh, by the grace of God King of England and of France, Lord of Ireland etc., I William Caxton, his most indyng, humble subject and little servant, PRESENT this little book unto the noble grace of my said lady; which book I late received in French from her good grace, and her commandment withal to reduce and translate it into our maternal and English tongue...

from the Prologue to Blanchardin and Eglantine, c. 1489

The evolutionary trajectory from Old to Middle English politeness can be seen as a change from an *IP* population (i) to an *AC* population intent on introducing politeness forms to praise the hearer, and (ii) to a *P* population where the latter forms are seen as more appropriate to address the hearer.

Finally, Leech describes the Modern English politeness as a process of growing democratization, understood as restricting deferential communicative behaviour. It led to a gradual reduction of the frequency in the use of honorifics and simplification of their structure (e.g. Noble Sir to Sir), until the recent trend of replacing titular nouns with first-name vocatives, often in an abbreviated form (e.g. Sue, Steve). We could interpret these changes as deflation of addressative politeness, which marks the onset of the *AQ* phase.

Needless to say, our interpretation of Leech's account grossly oversimplifies a lot of linguistic detail and a lot of linguistic facts. For example, a decline of honorific terms of address co-occurred with a growth of indirect directives, which goes against the analysis given above. Indirect requests having the interrogative form (Can you..., Will you...), which are standard nowadays, were infrequent in Old and Middle English, where the use of performative verbs (I command you, I beseech you) was favored. This shows, as noted before, that applying our model to actual populations must be exercised with due caution, particularly with regard to the type of linguistic data that are under consideration (e.g. terms of address vs. directives). But at the same time, we believe that the above account shows the utility of our model, which generates a set of predictions, which are testable and, hence, when applied to linguistic data can afford insight into politeness that goes beyond the unfalsifiable descriptivism of many works on politeness.

4.5 Future directions

Although politeness in conversation is part and parcel of the fabric of human interaction, and as such of close interest to social psychology, human ethology and human behavioral ecology (Brown 2015; Waciewicz et al. 2015), research into the social costs and benefits of speaking politely is in its infancy. Our analysis is a small first step in this direction, leaving numerous open questions that we see as fascinating directions for future research. One concerns the explanatory power of the constructs of perceived competence and warmth in the study of politeness, in particular the applicability of these constructs popular in research on professional settings (e.g. Belmi and

Pfeffer 2018) to wider contexts, as well as their ultimate-level bases. Another is the challenge of building more refined models, with more realistic strategies capable of capturing the differences in the distribution of perceived responsibility across different situation types, or agent-based modeling capable of capturing variations in individual valuations of competence vs. warmth. Finally, approaches from Social Exchange Theory describing transfers of material as well as non-material resources (cf. economic exchanges, Żywiczyński 2010) might provide insights necessary for addressing politeness in requests and offers.

Acknowledgements RM was supported by the Polish National Agency for Academic Exchange (NAWA) under grant agreement PPN/ULM/2019/1/00222. SW's work on this article was supported by the Polish National Agency for Academic Exchange (NAWA) under grant agreement PPN/BEK/2018/1/00382/U, as well as by the Polish National Science Centre (NCN) under grant agreement UMO-2019/34/E/HS2/00248.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix

A.1 Adaptation of extensive-form games from Chaudhry and Loewenstein

C&L model both situations as so-called extensive-form games. Figure 5 shows a version of each game²⁰, left the 'credit situation', and right the 'blame situation'. An extensive form game is a game-theoretic model that represents the sequential choices of different participants in a game, where the branches are terminal nodes that ascribe utility values to the players. Note that C&L choose different utility values for most outcomes with respect to w : while they argue that w changes for both players when one of both talks, we think that w can only change for the player that talks. This difference is stressed by setting the change of w in brackets whenever our model omits it. For more details about this difference, see Section A.2.

In the 'credit situation' between Roger (R) and Olivia (O), the former can first choose to thank her (T) or remain silent (Q). When Roger remains silent, then Olivia can choose to brag (B) or remain silent (Q). Depending of how both players behave, they gain or lose utility in form of perceived competence c or perceived warmth w , as represented at each branch, with the first value is Olivia's utility, the second is Roger's. In the blame situation, Olivia can choose between apologizing (A) and remaining silent (Q), and Roger between blaming (B) and remaining silent (Q). C&L show that thanking is the only rational choice in the credit situation and apologizing

²⁰ C&L reconsider multiple game versions for each situation, which differ in initial player, branch length and resulting payoffs. However, for the sake of demonstration it is sufficient to present one game for each situation, here each one with unspecified payoffs and branch length 2.

Fig. 5 Extensive form games of the credit situation (left) and the blame situation (right)

Table 10 Update of reputation in competence I_c and reputation in warmth I_w of originator O and receiver R for bragging, apologizing (where O is speaker), thanking, and blaming (where R is speaker)

	Bragging (O)	Thanking (R)	Apologizing (O)	Blaming (R)
$I_c(O)$	+	+	–	–
$I_w(O)$	–	+	+	–
$I_c(R)$	–	–	+	+
$I_w(R)$	–	+	+	–

‘+’ is an increase and ‘–’ a decrease of reputation

is the only rational choice in the blame situation²¹ for most settings of the parameter space.

A.2 Reputation updates in two different models

The difference between the update mechanisms in our model and C&L’s model is represented in Table 10 (C&L’s model) and Table 11 (our model). Note that in our model only the speaker’s reputation in warmth increases when she is polite and decreases when she is impolite. In the model of C&L both agents’ reputation in warmth increases when the speaker is polite and both decreases when the speaker is impolite. We believe that our update rule is more realistic, since we think that being polite or impolite can only have an impact on the speaker’s reputation in warmth and nobody else. The update in competence reputation is identical in both models.

A.3 The boosted RE (BRE) scenario

The boosted RE (BRE) scenario is a generalization of the RE scenario as given in Table 2, since it reconsiders the possibility that bragging and thanking together can have a stronger impact on the competence image c than only bragging or thanking alone. To

²¹ More in detail, C&L show that thanking/apologizing is the only so-called sub-game perfect Nash equilibrium (cf. Osborne 2004) for the respective game.

Table 11 Update of reputation in competence I_c and reputation in warmth I_w of originator O and receiver R for bragging, apologizing (where O is speaker), thanking, and blaming (where R is speaker)

	Bragging (O)	Thanking (R)	Apologizing (O)	Blaming (R)
$I_c(O)$	+	+	−	−
$I_w(O)$	−	0	+	0
$I_c(R)$	−	−	+	+
$I_w(R)$	0	+	0	−

‘+’ is an increase, ‘−’ a decrease, and ‘0’ no change of reputation

Table 12 The boosted responsibility exchange (BRE) scenario is a generalization of the RE scenario of Table 2, with the difference that the strategy combination B and T leads to a boosted exchange value of social image in competence c by factor β with $1 \leq \beta \leq 2$

	T	Q
B	$\beta c - w, w - \beta c$	$c - w, -c$
Q	$c, w - c$	$0, 0$

Table 13 The symmetric BRE scenario with generalized strategy labels

	AC	IP	P	AQ
AC	0	$(1 - \beta)c$	$(\beta - 1)c$	0
IP	$(\beta - 1)c - w$	$-w$	$\beta c - w$	$c - w$
P	$w - (\beta - 1)c$	$w - \beta c$	w	$w - c$
AQ	0	$-c$	c	0

model this aspect, we add an additional boost factor β by which c is reinforced for the strategy combination B, T , as given in Table 12.

We assume that $1 \leq \beta \leq 2$, so that for $\beta = 1$ we have the original scenario with no boost effect, and for $\beta = 2$ we have a doubling of the value, representing the case where bragging and thanking separately increases the competence image by c . The resulting symmetric table of the BRE scenario with generalized strategy labels is given in Table 13.

The evolutionary analysis of the BRE scenario reveals that P is the only ESS when $w > c$. However, when $c > w$ we have two different evolutionary outcomes depending on the magnitude of β . More concretely, when $\beta \leq \frac{w}{c} + 1$ then the outcome is the cycle as discussed in Sect. 3.3. When $\beta > \frac{w}{c} + 1$ then the strategy IP is the only ESS. The different conditions and the respective results are depicted in Table 14.

The case where IP is an ESS constitutes a situation similar to the famous prisoners dilemma (PD) game. In the PD game, mutual defection results in a worse outcome for anybody than mutual cooperation, but mutual defection is the only (strict) NE and ESS, since no matter what the other player would do, it is always rational (since utility-maximizing) to play D . For the case where $c > w$ and $\beta > \frac{w}{c} + 1$, Table 13 entails the same dilemma. Playing mutually IP results in $-w$ and is worse than any other mutual profile (mutual AC and AQ results in 0, mutual P in w), but no matter

Table 14 Results of the analysis of the boosted responsibility exchange (BRE) scenario: when $w > c$ the P is the only ESS of the game

	$w > c$	$c > w$
$\beta \leq \frac{w}{c} + 1$	P	cycle
$\beta > \frac{w}{c} + 1$	P	IP

When $c > w$ then there are two cases. When $\beta \leq \frac{w}{c} + 1$ then the result under evolutionary dynamics is a cyclic process as described in Sect. 3.3. When $\beta > \frac{w}{c} + 1$, IP is the only ESS of the game

what the other player would do, it is always rational (since utility-maximizing) to play IP .

A.4 The symmetric RE scenario with bother player's utility values

The symmetric RE scenario utility table with both player's utility values is given in Table 15.

A.5 Average population shares over time

Let $f(s, t)$ be the population share of strategy s at time t . Then the average population share f_a over the past at time t can be defined as

$$f_a(s, t) = \frac{1}{t} \int_{\tau=0}^t f(s, \tau) d\tau$$

Figure 6 shows the change of average population shares over time under replicator dynamics for the game of Table 4 with $c = 2w$ and starting population: $f(AC) = f(IP) = 0.24$ and $f(AQ) = f(P) = 0.26$. As observable, all four values oscillate around 0.25 with the same increase in amplitude, what indicated that all strategies an average population shares f_a of 0.25 when $t \rightarrow \infty$.

A.6 Population shares over time for different w - c relationships

Table 15 The symmetric RE scenario defined as utility table over strategy pairs of originator and receiver strategy with row player's utility (first value) and column player's utility (second value)

	$\langle B, T \rangle$	$\langle B, Q \rangle$	$\langle Q, T \rangle$	$\langle Q, Q \rangle$
$\langle B, T \rangle$	0, 0	0, $-w$	0, w	0, 0
$\langle B, Q \rangle$	$-w$, 0	$-w$, $-w$	$c - w$, $w - c$	$c - w$, $-c$
$\langle Q, T \rangle$	w , 0	$w - c$, $c - w$	w , w	$w - c$, c
$\langle Q, Q \rangle$	0, 0	$-c$, $c - w$	c , $w - c$	0, 0

Note that the column player's utilities corresponds to the row player's utilities by mirroring at the top-left to bottom-right diagonal

Figure 7 shows the change in population shares for $c = \frac{3}{2}w$ and Fig. 8 for $c = 3w$ (initial population as in Fig. 3). In the first case, the strategy AQ is dominant at the

Fig. 6 Changes of average population shares over time under replicator dynamics for the game of Table 4 with $c = 2w$

Fig. 7 Changes of population shares over time under replicator dynamics for the game of Table 4 with $c < 2w$ (concretely: $c = \frac{3}{2}w$)

Fig. 8 Changes of population shares over time under replicator dynamics for the game of Table 4 with $c > 2w$ (concretely: $c = 3w$)

Table 16 The symmetric OS scenario's situations defined as utility table over strategy pairs of originator and receiver strategy and deduced from the (asymmetric) OS scenario's situations of Table 7

	$\langle B, T \rangle$	$\langle B, Q \rangle$	$\langle Q, T \rangle$	$\langle Q, Q \rangle$
(a) <i>Situation 1: good service</i>				
$\langle B, T \rangle$	c	c	c	c
$\langle B, Q \rangle$	$c - w$	$c - w$	$c - w$	$c - w$
$\langle Q, T \rangle$	$c + w$	w	$c + w$	w
$\langle Q, Q \rangle$	c	0	c	0
	$\langle A, C \rangle$	$\langle A, Q \rangle$	$\langle Q, C \rangle$	$\langle Q, Q \rangle$
(b) <i>Situation 2: bad service</i>				
$\langle A, C \rangle$	$-c$	$-c$	$-c$	$-c$
$\langle A, Q \rangle$	$w - c$	$w - c$	$w - c$	$w - c$
$\langle Q, C \rangle$	$-w - c$	$-w$	$-w - c$	$-w$
$\langle Q, Q \rangle$	$-c$	0	$-c$	0

expense of strategy AC , whereas in the second case, the strategy AC is dominant at the expense of strategy AQ . In both cases the strategies P and IP get the same amount of population share over time.

A.7 Evolutionary analysis of the OS scenario

Table 16(a) shows the symmetric utility table over strategy pairs for the good service scenario. Note that the polite strategy P , represented by the strategy pair $\langle Q, T \rangle$, achieves a utility value of $c + w$ against itself. Since $c, w > 0$, it is easy to see that this is the unique maximum in the third column, independent of $c < w$ or $c > w$. Thus, the strategy is a strict Nash equilibrium and therefore an evolutionary stable strategy. Furthermore, it is also easy to spot that the other three strategies pairs' utility value against itself is not maximal, independent of $c < w$ or $c > w$. Thus, none of them is evolutionarily stable.

Table 16(b) shows the symmetric utility table over strategy pairs for the bad service scenario. Note that the polite strategy P , here represented by the strategy pair $\langle A, Q \rangle$, achieves a utility value of $w - c$ against itself. It is only the unique maximum if its column, and therefore a strict NE and an ESS, iff $w > c$. Furthermore, in this case it is the only ESS. However, in case $c > w$, the utility of strategy pair $\langle Q, Q \rangle$ against itself is a unique maximum of its column, and therefore a strict NE and an ESS, in fact the only ESS of the game table.

References

- Agha, A. (1994). Honorification. *Annual Review of Anthropology*, 23, 277–302.
- Ahern, C. A., & Quinley, J. (2014). Would you trust me please? A formal model of polite requests and relationships. Paper presented at the 19th International Congress of Linguists, Geneva, 2013. https://www.cil19.org/uploads/documents/Would_you_trust_me_please.pdf.

- Asher, N., & Quinley, J. (2012). Begging questions, their answers and basic cooperativity. In M. Okumura, D. Bekki, & K. Satoh (Eds.), *New frontiers in artificial intelligence* (LNCS 7258, pp. 3–12). Heidelberg: Springer.
- Belmi, P., & Pfeffer, J. (2018). The effect of economic consequences on social judgment and choice: Reward interdependence and the preference for sociability versus competence. *Journal of Organizational Behavior*, 39(8), 990–1007.
- Blum-Kulka, S. (1992). The metapragmatics of politeness in Israeli society. In R. Watts, S. Ide, & K. Ehlich (Eds.), *Politeness in language: Studies in its history, theory and practice* (pp. 255–280). Berlin: Mouton de Gruyter.
- Brambilla, M., Sacchi, S., Castellini, F., & Riva, P. (2010). The effects of status on perceived warmth and competence: Malleability of the relationship between status and stereotype content. *Social Psychology*, 41, 82–87.
- Brinöl, P., & Petty, R. E. (2009). Source factors in persuasion: A self-validation approach. *European Review of Social Psychology*, 20, 49–96.
- Brown, P. (2015). Politeness and language. In J. D. Wright (Ed.), *International encyclopedia of the social and behavioral sciences* (2nd ed., pp. 326–330). Amsterdam: Elsevier.
- Brown, P. (2017). Politeness and impoliteness. In Y. Huang (Ed.), *Oxford handbook of pragmatics* (pp. 383–399). Oxford: Oxford University Press.
- Brown, P., & Levinson, S. (1978). Universals in language usage: Politeness phenomena. In E. Goody (Ed.), *Questions and politeness: Strategies in social interaction* (pp. 56–311). Cambridge: Cambridge University Press.
- Brown, P., & Levinson, S. (1987). *Politeness: Some universals in language usage*. Cambridge: Cambridge University Press.
- Brown, R., & Ford, M. (1961). Address in American English. *The Journal of Abnormal and Social Psychology*, 62(2), 375–385.
- Brown, R., & Gilman, A. (1960). The pronouns of power and solidarity. In T. Sebeok (Ed.), *Style in language* (pp. 253–276). Cambridge, MA: MIT Press.
- Chaudhry, S., & Loewenstein, G. (2019). Thanking, apologizing, bragging, and blaming: Responsibility exchange theory and the currency of communication. *Psychological Review*, 126(3), 313–344.
- Clark, R. (2012). *Meaningful games: Exploring language with game theory*. Cambridge, MA: MIT Press.
- Culpeper, J. (2011). *Impoliteness using language to cause offence*. Cambridge: Cambridge University Press.
- Danescu-Niculescu-Mizil, C., Sudhof, M., Jurafsky, D., Leskovec, J., & Potts, C. (2013). A computational approach to politeness with application to social factors. [arxiv:1306.6078](https://arxiv.org/abs/1306.6078).
- Eelen, G. (2001). *A critique of politeness theories*. Manchester: St. Jerome.
- Fehr, E., Fischbacher, U., & Gächter, S. (2002). Strong reciprocity, human cooperation, and the enforcement of social norms. *Human Nature*, 13, 1–25.
- Fiske, A. (1992). The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychological Review*, 99, 689–723.
- Fiske, A. (2004). Four modes of constituting relationships: Consubstantial assimilation; space, magnitude, time, and force; concrete procedures; abstract symbolism. In N. Haslam (Ed.), *Relational models theory: A contemporary overview* (pp. 61–146). Mahwah, NJ: Lawrence Erlbaum.
- Floyd, S., Rossi, G., Baranova, J., Blythe, J., Dingemanse, M., Kendrick, K. H., et al. (2018). Universals and cultural diversity in the expression of gratitude. *Royal Society Open Science*, 5(5), 180391.
- Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. *Zeitschrift für Sprachwissenschaft*, 35(1), 3–44.
- Giardini, F., Fitneva, S. A., & Tamm, A. (2019). “Someone told me”: Preemptive reputation protection in communication. *PloS ONE*, 14(4), e0200883.
- Gintis, H. (2000). Strong reciprocity and human sociality. *Journal of Theoretical Biology*, 206, 169–179.
- Goffman, E. (1967). *Interaction ritual: Essays on face-to-face behavior*. New York: Doubleday.
- Grice, H. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics 3: Speech acts* (pp. 41–58). New York: Academic Press.
- Henrich, J., & Gil-White, F. J. (2001). The evolution of prestige: Freely conferred deference as a mechanism for enhancing the benefits of cultural transmission. *Evolution and Human Behavior*, 22(2), 165–196.
- Jucker, A. (2012). Changes in politeness cultures. In E. C. Traugott & T. Nevalaine (Eds.), *The Oxford handbook of the history of English*. New York: Oxford University Press.
- Kádár, D. Z., & Haugh, M. (2013). *Understanding politeness*. Cambridge, MA: Oxford University Press.

- Kutsukake, N., & Hosaka, K. (2015). Aggression and conflict management. In M. Nakamura, K. Hosaka, N. Itoh, & K. Zamma (Eds.), *Mahale chimpanzees: 50 years of research* (pp. 448–458). Cambridge: Cambridge University Press.
- Lakoff, R. (1973). The logic of politeness, or minding your P's and Q's. In C. Corum, T. C. Smith-Stark, & A. Weiser (Eds.), *Papers from the 9th regional meeting of the Chicago Linguistic Society* (pp. 292–305). Chicago, IL: Chicago Linguistic Society.
- Lakoff, R. (1990). *Talking power: The politics of language in our lives*. Glasgow: Harper Collins.
- Lee, J. J., & Pinker, S. (2010). Rationales for indirect speech: The theory of the strategic speaker. *Psychological Review*, 117(3), 785–807.
- Leech, G. (1983). *The principles of pragmatics*. London: Longman.
- Leech, G. (2014). *The pragmatics of politeness*. Oxford: Oxford University Press.
- Lee-Wong, S. (1999). *Politeness and face in Chinese culture*. Frankfurt: Peter Lang.
- Locher, M. (2012). Politeness. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. New York: Wiley-Blackwell. <https://doi.org/10.1002/9781405198431.wbeal0916>.
- Marčjanik, M. (2015). Współczesna etykieta językowa-zmierzch wartości? *Studia Medioznawcze*, 63(4), 115–119.
- Maynard Smith, J., & Price, G. R. (1973). The logic of animal conflict. *Nature*, 246(5427), 15–18.
- McCready, E. (2015). *Reliability in pragmatics*. Oxford: Oxford University Press.
- McCready, E. (2019). *The semantics and pragmatics of honorification*. Oxford: Oxford University Press.
- Mühlenbernd, R., Waciewicz, S., & Żywiczyński, P. (2019). The game theory of politeness in language: A formal model of polite requests. <https://doi.org/10.31234/osf.io/6srux>
- Myerson, R. B. (1991). *Game theory: Analysis of conflict*. Cambridge, MA: Harvard University Press.
- Noë, R., & Hammerstein, P. (1995). Biological markets. *Trends in Ecology and Evolution*, 10(8), 336–339.
- Nowak, M. A. (2006). Five rules for the evolution of cooperation. *Science*, 314, 1560–1563.
- Ohtsuki, H., & Iwasa, Y. (2006). The leading eight: Social norms that can maintain cooperation by indirect reciprocity. *Journal of Theoretical Biology*, 239(4), 435–444.
- Oosterbeek, H., Sloof, R., & van de Kuilen, G. (2004). Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2), 171–188.
- Osborne, M. J. (2004). *An introduction to game theory*. Oxford: Oxford University Press.
- Östman, J.-O. (1989). Testing iconicity: Sentence structure and politeness. *Belgian Journal of Linguistics*, 4, 145–163.
- Pinker, S. (2007). *The stuff of thought: Language as a window into human nature*. New York: Viking.
- Pinker, S., Nowak, M., & Lee, J. (2008). The logic of indirect speech. *Proceedings of the National Academy of Sciences*, 105(3), 833–838.
- Quinley, J. (2012). Trust games as a modal request. In D. Lassiter & M. Slavkovik (Eds.), *New directions in logic, language and computation* (LNCS 7415, pp. 221–233). Heidelberg: Springer.
- Quinley, J. & Ahern, C. (2012). Question of trust. In S. Katrenko & R. K. Rendsvig (Eds.), *Proceedings of the 2012 ESSLI student session* (pp. 121–132). Heidelberg: CEUR Workshop Proceedings.
- Sakamaki, T., & Hayaki, H. (2015). Greetings and dominance. In M. Nakamura, K. Hosaka, N. Itoh, & K. Zamma (Eds.), *Mahale chimpanzees: 50 years of research* (pp. 459–471). Cambridge: Cambridge University Press.
- Taylor, P. D., & Jonker, L. B. (1978). Evolutionarily stable strategies and game dynamics. *Mathematical Biosciences*, 40, 145–156.
- Tomasello, M. (2008). *Origins of human communication*. Cambridge, MA: MIT Press.
- Torreira, F., Bögels, S., & Levinson, S. C. (2015). Breathing for answering: The time course of response planning in conversation. *Frontiers in Psychology*, 6, 284.
- van Rooy, R. (2003). Being polite is a handicap: Towards a game theoretical analysis of polite linguistic behavior. In *TARK'03: Proceedings of the 9th conference on Theoretical Aspects of Rationality and Knowledge* (pp. 45–58). New York: ACM.
- Waciewicz, S., Żywiczyński, P., & McCrohon, L. (2015). Linguistic politeness from an ethological perspective: Theoretical questions and empirical issues. *Theoria et Historia Scientiarum*, 11, 81–98.
- Watts, R. (2003). *Politeness*. Cambridge: Cambridge University Press.
- Zahavi, A., & Zahavi, A. (1997). *The handicap principle: A missing piece of Darwin's puzzle*. New York: Oxford University Press.
- Żywiczyński, P. (2010). The axiology of spoken interaction: An essay on the organisation of conversational behaviour. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika.

Żywczyński, P. (2012). An attempt at a proxemic description of politeness from the ethological-evolutionary perspective. *Theoria et Historia Scientiarum*, 9, 71–98.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.