
Signaling games and non-literal meaning

MERLIJN SEVENSTER

Institute for Logic Language and Computation, Universiteit van Amsterdam

sevenstr@science.uva.nl

ABSTRACT. Sally (Sally 2003) proposed a game-theoretical framework to study the use and conventionalization of non-literal language use. We have objections against this model from a conceptual and a methodological point of view. Instead, we introduce a ‘*Super Conventional signaling game*’ to account for non-literal use of language. Such a game is an elaboration of Lewis’ (Lewis 1969) ‘signaling game’, that preserves and makes use of a strict separation between literal and non-literal interpretation of utterances. We argue that, following Sally now, real people play Super Conventional signaling games as they use non-literal speech.

1 Introduction

¹ If one says “George W. Bush is a pig”, one does not mean to say that George W. Bush is a pink animal with a tail, although this is implied by the literal meaning of the sentence. Instead, the speaker of this sentence expects the hearer to give the sentence its non-literal meaning. Sally (Sally 2003) recognizes the use of messages that are intended to have a non-literal meaning as ‘risky speech’, as the hearer might not recognize the speaker’s intention and respond “Oh, that comes as a surprise to me. I thought he was the president of the US”. Although more risky, experimental results (for references see (Sally 2003)) show that non-literal speech has stronger cognitive and social impact than literal speech.

Lewis (Lewis 1969) defined the notion of a signaling game in order to explain the conventionalization of meaning of language without assuming any pre-existing relation between messages and meaning. Playing a signaling game optimally (in a sense that will be made explicit below) results in a unique meaning for every message. Lewis used signaling games as a philosophical reply to Quine’s objection against conventionalism, that any linguistic convention could only be formed using another linguistic convention. The last couple of years however, scholars use signaling games to come to an understanding of pragmatical issues; e.g. Horn’s division of pragmatic labor (van Rooij 2004).

A signaling game is a cooperative game amongst two players: a sender **S** and a receiver **R**, whose shared goal it is to let **R** perform an action that is appropriate with respect to

¹We gratefully acknowledge Robert van Rooij for guiding remarks, his continuing enthusiasm and for proof-reading this paper several times. We thank Peter van Emde Boas for correcting big mistakes. We thank the anonymous referees for numerous hints to improve the paper.

the state S and R are in. This state is only observed by S though, and S communicates it by means of a meaningless message. Think of a couple of agents, the one looking out for hungry predators and the other searching for food on the ground. Both players are in the same state — there is a predator approaching or there is not —, but only S knows which state they are in. S signals the state by means of a message that has no pre-defined meaning — say, ‘buh’ or ‘bah’. In turn, R hears the message and is free to perform any action of his liking, but every state has a most appropriate action. E.g. if there is a hungry predator approaching, R should flee, otherwise R should keep on searching. The best thing for S to do is to say ‘buh’ if there is a hunting predator and ‘bah’ otherwise; and for R to flee when he hears ‘buh’ and not to in case he hears ‘bah’. (It is equally good of course to do the same, but with ‘buh’ and ‘bah’ interchanged.) If S and R play the game in such an optimal way, the meanings of the messages ‘buh’ and ‘bah’ are created in the play of the game. Game-theoretically, playing in a way that makes the messages meaningful amounts to coordinating on a ‘*separating Nash equilibrium*’ (to be defined below).

In this paper, we are concerned with conventionalization of *non-literal* meaning. As to non-literal meaning, we distinguish, following Searle (Searle 1979), between what a sentence means and what a speaker means by uttering this sentence. The non-literal meaning of a sentence is its ‘speaker’s utterance meaning’.² In the hearer’s process of attaching the speaker’s utterance meaning — according to Searle, and we follow him —, the hearer first has to recognize the defectiveness of the utterance’s literal meaning. For this recognition to be possible, it thus must be the case that the literal meaning is common knowledge. This insight we mark (♠). As to the conventionalization of non-literal meaning, we propose a signaling game à la Lewis that takes the conventional meaning as parameter to model conventionalization of *non-literal meaning*; a proposal that is conceptually different from the one proposed by Sally (Sally 2003). In his ‘*general signaling games*’ namely, Sally does not take the state of the players as primitive, but rather the speech act assigned to S . We feel that this does not do right to the notion of speech act, as it neglects the performative side of playing signaling games. We will come back to Sally’s model, after we have defined Lewisian signaling games and discuss our objections with respect to this model.

Although we conceptually disagree with Sally’s signaling game, we are more than happy to adopt his insights on how people play signaling games in experimental settings and how these findings can be used to explain under what conditions people are more likely to use non-literal speech.

Sally namely, taking for granted that “people play the language game in a way that is consistent with their play in all games” ((Sally 2003), pg. 1232), introduces empirically justified rules of thumb that describe how people play coordination games. The central notion in these rules appears to be ‘sympathy’ and ‘common ground’, as the rules roughly state that players are more likely to play risky if they are more sympathetic to each other.

The element of risk is introduced by Sally’s pay-off functions that reward successful non-literal communication higher than successful literal communication; and that — in case

²We are afraid that this is closest we will get to defining the linguistic notion of ‘non-literal meaning’. Eventually, we will give a formal definition.

of unsuccessful communication — punish the player (possibly both) deviating from the convention more severely than if he would have stuck to the convention. For these pay-off functions empirical justification is given, but they are not defined technically. Sally shows that playing his game riskily amounts to using messages non-literally. As a consequence it follows that the more sympathetic people are, the more likely it is for them to communicate non-literally: A finding that itself matches our intuition and empirical reports.

In our approach, we take the notion of literal meaning for granted and use it in a higher-level signaling game to account for conventional, non-literal meaning. We call this game a ‘*Super Conventional signaling game*’ (from now on SC signaling game), being a game that has a convention of literal meaning parameterized. We will prove that also in our Super Conventional signaling games playing (more) risky resembles communicating (more) non-literally.

In Section 2 we formally define signaling games in strategic normal form and introduce the game theoretical solution concepts of pay-off dominant and risk dominant Nash equilibria. The latter will be used to formalize the notion of meaning, the former to formalize the notion of playing risky. In Section 3 we will introduce Sally’s model in greater length, point out our objections, and state Sally’s rules of thumb as to how game players actually play coordination games and see how these are used to explain under what conditions people use non-literal speech. In Section 4 we will define the notion of a Super Conventional signaling game. We will prove that this game has multiple pay-off dominant and one risk dominant Nash equilibrium, using parameter values inspired by Sally’s findings. We will conclude by showing that also in Super Conventional signaling players play more risky, if they are more sympathetic. Section 5 fulfills its role as a conclusion.

2 Signaling games

Let T be the set of states, M be the set of messages and A be the set of actions such that $|T| = |A| \leq |M|$. Let $f : T \rightarrow A$ be the bijective function, that adds the appropriate action $f(t) \in A$ to every type $t \in T$. Then \mathbf{S} (\mathbf{R}) plays the signaling game following strategy \mathbf{s} (\mathbf{r}), that is a function from $T \rightarrow M$ ($M \rightarrow A$). In *cheap talk signaling games*, successful communication of state t (thus in case $f(t) = \mathbf{r}(\mathbf{s}(t))$) is rewarded with 1, whereas unsuccessful communication (thus in case $f(t) \neq \mathbf{r}(\mathbf{s}(t))$) is rewarded with 0, independent of the state t and the message $\mathbf{s}(t)$:

$$u_{\mathbf{S}}(f, t, \mathbf{s}, \mathbf{r}) = u_{\mathbf{R}}(f, t, \mathbf{s}, \mathbf{r}) = \begin{cases} 1, & \text{if } f(t) = \mathbf{r}(\mathbf{s}(t)); \\ 0, & \text{if } f(t) \neq \mathbf{r}(\mathbf{s}(t)). \end{cases} \quad (1.1)$$

We assume that Nature picks the state according to some probability distribution \mathbf{P} over T .³ The utility function for \mathbf{S} is the expected utility relative to the probability distribution \mathbf{P} over T :

³We assume that $\mathbf{P}(t) > 0$, for every $t \in T$.

$$U_S(\mathbf{s}, \mathbf{r}) = \sum_{t \in T} \mathbf{P}(t) u_S(f, t, \mathbf{s}, \mathbf{r}) \quad (1.2)$$

and the utility function for \mathbf{R} is the expected utility provided that \mathbf{S} uses \mathbf{s} , so that \mathbf{S} is of a type from $T_{\mathbf{s}(t)} = \{t' \in T \mid \mathbf{s}(t) = \mathbf{s}(t')\}$:

$$U_R(\mathbf{s}, \mathbf{r}) = \sum_{t \in T} \mathbf{P}(t) \sum_{t' \in T} \mathbf{P}(t' \mid T_{\mathbf{s}(t)}) u_R(f, t', \mathbf{s}, \mathbf{r}). \quad (1.3)$$

Finally, we define a *cheap talk signaling game* \mathbf{G} as a tuple $\langle \{\mathbf{S}, \mathbf{R}\}, \{\mathbf{S}, \mathbf{R}\}, \{U_S, U_R\} \rangle$, where \mathbf{S} is the set of strategies $\mathbf{s} : T \rightarrow M$ for player \mathbf{S} ; \mathbf{R} is the set of strategies $\mathbf{r} : M \rightarrow A$ for player \mathbf{R} ; and $\{U_S, U_R\}$ contains both players' utility functions as defined in (1.2) and (1.3). \mathbf{G} is called 'cheap talk' because u_S and u_R , simultaneously defined in (1.1) are called this way. Our SC signaling game will use other functions.

A pair of strategies $\langle \mathbf{s}^*, \mathbf{r}^* \rangle$ forms a *Nash equilibrium*⁴ in \mathbf{G} iff neither \mathbf{S} nor \mathbf{R} gains from unilateral deviation:

$$U_S(\mathbf{s}^*, \mathbf{r}^*) \geq U_S(\mathbf{s}, \mathbf{r}^*) \quad \text{and} \quad U_R(\mathbf{s}^*, \mathbf{r}^*) \geq U_R(\mathbf{s}^*, \mathbf{r}),$$

for all $\mathbf{s} \in \mathbf{S}$ and $\mathbf{r} \in \mathbf{R}$. We will omit to mention the game \mathbf{G} , if no confusion arises. We call a Nash equilibrium $\langle \mathbf{s}^*, \mathbf{r}^* \rangle$ *pay-off dominant* in \mathbf{G} iff $U_S(\mathbf{s}^*, \mathbf{r}^*) \geq U_S(\mathbf{s}, \mathbf{r})$ and $U_R(\mathbf{s}^*, \mathbf{r}^*) \geq U_R(\mathbf{s}, \mathbf{r})$, for all Nash equilibria $\langle \mathbf{s}, \mathbf{r} \rangle$ in \mathbf{G} . Conversely, we say that $\langle \mathbf{s}, \mathbf{r} \rangle$ is *pay-off dominated* by $\langle \mathbf{s}^*, \mathbf{r}^* \rangle$. To strengthen our intuition, consider the following utility matrix:

	\mathbf{r}_1	\mathbf{r}_2
\mathbf{s}_1	2, 2	2, 0
\mathbf{s}_2	0, 2	3, 3

Strategy pairs $\langle \mathbf{s}_1, \mathbf{r}_1 \rangle$ and $\langle \mathbf{s}_2, \mathbf{r}_2 \rangle$ are Nash equilibria; and $\langle \mathbf{s}_2, \mathbf{r}_2 \rangle$ is the payoff dominant equilibrium. Although $\langle \mathbf{s}_1, \mathbf{r}_1 \rangle$ is pay-off dominated, \mathbf{S} and \mathbf{R} have their incentive to choose \mathbf{s}_1 and \mathbf{r}_1 , respectively: both strategies guarantee a utility of 2 and are therefore considered less risky. Formally, we call a Nash equilibrium $\langle \mathbf{s}^*, \mathbf{r}^* \rangle$ *risk dominant* in \mathbf{G} in the sense of (Harsanyi and Selten 1988) iff

$$\begin{aligned} & (U_S(\mathbf{s}^*, \mathbf{r}^*) - U_S(\mathbf{s}, \mathbf{r}^*))(U_R(\mathbf{s}^*, \mathbf{r}^*) - U_R(\mathbf{s}^*, \mathbf{r})) \\ & \geq \\ & (U_S(\mathbf{s}, \mathbf{r}) - U_S(\mathbf{s}^*, \mathbf{r}))(U_R(\mathbf{s}, \mathbf{r}) - U_R(\mathbf{s}, \mathbf{r}^*)), \end{aligned}$$

for all Nash equilibria $\langle \mathbf{s}, \mathbf{r} \rangle$ in \mathbf{G} . As the reader can check, $\langle \mathbf{s}_1, \mathbf{r}_1 \rangle$ is risk dominant.

For instance, consider a signaling game where $T = \{t_1, t_2\}$, $M = \{m_1, m_2\}$ and $A = \{a_1, a_2\}$ and $x = \mathbf{P}(t_1) > \mathbf{P}(t_2) = y$. Then \mathbf{S} and \mathbf{R} both have four different strategies, hence there are 16 strategy-pairs of which we have the players' utilities U_S and U_R below.

⁴In fact, we are dealing with a *Bayesian* Nash equilibrium. For definitions consult (Osborne and Rubinstein 1994).

	t_1	t_2		m_1	m_2				
\mathbf{s}_1	m_1	m_1	\mathbf{r}_1	a_1	a_1				
\mathbf{s}_2	m_1	m_2	\mathbf{r}_2	a_1	a_2				
\mathbf{s}_3	m_2	m_1	\mathbf{r}_3	a_2	a_1				
\mathbf{s}_4	m_2	m_2	\mathbf{r}_4	a_2	a_2				

	\mathbf{r}_1	\mathbf{r}_2	\mathbf{r}_3	\mathbf{r}_4
\mathbf{s}_1	x, x	x, x	y, y	y, y
\mathbf{s}_2	x, x	$1, 1$	$0, 0$	y, y
\mathbf{s}_3	x, x	$0, 0$	$1, 1$	y, y
\mathbf{s}_4	x, x	x, x	y, y	y, y

This cheap talk signaling game has four Nash equilibria, being $\langle \mathbf{s}_1, \mathbf{r}_1 \rangle$, $\langle \mathbf{s}_2, \mathbf{r}_2 \rangle$, $\langle \mathbf{s}_3, \mathbf{r}_3 \rangle$ and $\langle \mathbf{s}_4, \mathbf{r}_1 \rangle$. As the reader can check, only in $\langle \mathbf{s}_2, \mathbf{r}_2 \rangle$ and $\langle \mathbf{s}_3, \mathbf{r}_3 \rangle$ communication takes place: these are precisely the pay-off dominant equilibria. Lewis calls such equilibria ‘signaling systems’. Technically, $\langle \mathbf{s}, \mathbf{r} \rangle$ is a signaling system iff $f(t) = \mathbf{r}(\mathbf{s}(t))$, for every $t \in T$. Necessary condition for $\langle \mathbf{s}, \mathbf{r} \rangle$ be a signaling system is that \mathbf{s} and \mathbf{r} are bijective functions.

3 Sally’s model and four rules of thumb

Sally gives a sketchy account of a signaling game that models the use of non-literal speech (and conventionalization of non-literal meaning). Contrary to Lewis, Sally takes speech acts (SA) as primitives rather than states. So Nature assigns \mathbf{S} a speech act $sa \in SA$, and it is up to \mathbf{R} to interpret \mathbf{S} ’s message $m \in M$ as the intended speech act sa . Sally (Sally 2003) defines a ‘general signaling game’ \mathbf{G}' as a tuple $\langle \{\mathbf{S}, \mathbf{R}\}, \{\mathbf{S}', \mathbf{R}'\}, \{U'_S, U'_R\} \rangle$, where \mathbf{S} and \mathbf{R} are the two players as we had them before; \mathbf{S}' (\mathbf{R}') is the set of strategies $SA \rightarrow M$ ($M \rightarrow SA$) for player \mathbf{S} (\mathbf{R}); and the utility function U'_S and U'_R capture “the setting, the needs and the wants of the participants, their relationship [and] their prior statements” ((Sally 2003), pg. 1232). Sally (Sally) has shown that every strategy-pair $\langle \mathbf{s}', \mathbf{r}' \rangle$ such that $sa = \mathbf{r}'(\mathbf{s}'(sa))$ for all $sa \in SA$ is a Nash equilibrium.

Other than in cheap talk signaling games two separating Nash equilibria $\langle \mathbf{s}', \mathbf{r}' \rangle$ and $\langle \mathbf{s}'', \mathbf{r}'' \rangle$ do not yield equivalent utilities per se; i.e. possibly $U'_S(\mathbf{s}', \mathbf{r}') \neq U'_S(\mathbf{s}'', \mathbf{r}'')$ and $U'_R(\mathbf{s}', \mathbf{r}') \neq U'_R(\mathbf{s}'', \mathbf{r}'')$. This differing pay-off models the differing cognitive and social impact⁵ of the messages used to communicate a certain speech act. In particular, Sally gives empirical evidence for the fact that successful communication by means of a non-literally intended message has greater impact than a literally intended message. On the other hand, in the case of unsuccessful communication \mathbf{S} sent a non-literally intended message and/or \mathbf{R} interpreted the received message non-literally. Sally gives empirical evidence that this situation is more painful for the player (possible both) who deviated from the literal meaning. Moreover, Sally claims that the penalty for the player deviating from the literal meaning in case of unsuccessful communication is greater than the reward the player experiences in case of successful non-literal communication. Intuitively we agree with this: One can call one’s friend a wanker, but how extremely embarrassing is the conversation wherein the supposed friend takes it as an insult!

⁵People have shown to remember non-literal utterances better than literal ones. Furthermore, use of non-literal speech implicitly stresses the common ground of the interlocutors. For references see (Sally 2003).

Our criticism to Sally’s general signaling game is twofold. Firstly, by taking the speech acts as primitive objects, Sally neglects the performative aspect of signaling games. It is in Lewisian signaling games the message $\mathbf{s}(t)$ together the act of playing the game according to \mathbf{s} that causes $\mathbf{s}(t)$ become a speech act in itself. As such, Lewisian signaling games have more atomic primitives than general signaling games. Secondly, we sense that pay-off functions U'_S and U'_R in their current form take so many parameters into account that any formal elaboration is doomed to fail. Furthermore, the utility-functions do not reflect any philosophical considerations as to how non-literal language is opposed to literal language. E.g. is the notion of literal language incorporated in the “setting” or the interlocutors’ “prior statements”? In the next section we will rigorously define our signaling games, that take the states as primitives and have the literal meaning of the message parameterized.

On the other hand, we will adopt Sally’s insights under what condition people tend to communicate non-literally. Sally namely takes for granted that “people play the language game in a way that is consistent with their play in all games”. We think that he is right in doing so, and like to quote Rubinstein in favor of this stance: “[...] if game theory is to shed light on real life phenomena, linguistic phenomena are the most promising candidates. Game theoretical solution concepts are most suited to stable life situations which are “played” often by large populations of players”. Hence, the condition that causes players to coordinate on such-and-such an equilibrium also causes interlocutors to communicate in a way that corresponds to that type of equilibrium. In particular, Sally links pay-off dominant Nash equilibria to a convention of non-literal meaning that is more risky than the convention of literal meaning.

Below we have the rules stated as to how people play coordination games and as to under what conditions interlocutors communicate non-literally. We will use them in the same way Sally does.

Rules of thumb – Coordination games Although Harsanyi and Selten (Harsanyi and Selten 1988) argued otherwise, people have been shown to coordinate on the risk dominant Nash equilibria by default. Sally summarized these results ((Sally 2003), pg. 1229–1231, for references to empirical evidence (Sally 2003)):

Rule 1: “In a game with one outcome risk dominant and another ‘modestly’ pay-off dominant, the former is more likely to be chosen.”

Rule 2: “As sympathy between the players increases, a pay-off dominant, risk dominated equilibrium is more likely to be realized.”

Rules of thumb – Non-literal language usage In the literature it is argued that non-literal speech can only be used if the common ground of speaker and hearer is big enough. For otherwise the receiver can not make up whether the speaker intended the utterances literally or metaphorically. We summarize the results as follows (for references to empirical evidence see (Sally 2003)):

Rule i: “If the common ground is minimal, people are more likely to speak literally than non-literally.” We take this rule for granted.

Rule ii: “The more the common ground of interlocutors come to overlap, the more they will use non-literal language.”

It is not too daring an assumption that the more people are sympathetic to each other, the more their common ground overlaps. In the next section, we will define the notion of a SC signaling game. In Section 5 we will predict which equilibrium will be satisfied under what conditions assuming that players play this game following Rules 1 and 2.

4 Super Conventional Signaling Games

Intuition behind SC signaling games says that **S** and **R** play a signaling game, having common knowledge of the fact that $\langle \mathbf{cs}, \mathbf{cr} \rangle$ is the conventional signaling system. Thus, we denote the conventional sender and receiver strategy by means of \mathbf{cs} and \mathbf{cr} , respectively. It is these strategies that model the literal meaning. As such they have agreed on the conventional meaning of messages in

$$M' = \{m \in M \mid \text{there exists a } t \in T \text{ such that } \mathbf{s}(t) = m\},$$

that contains the messages that convey the to-be communicated types. Since \mathbf{s} is a function, $|M'| = |T|$. In accordance with Searle (Searle 1979), it are only the messages in M' that can have non-literal meaning.

Typically, non-literal utterances have it that the sentence taken literally means something different than was intended by the speaker. In formal terms, although **S** is of type t he uses a message $m \neq \mathbf{cs}(t)$, and **S** wants and expects **R** not to perform $\mathbf{cr}(m)$ but $f(t)$.

Following Sally we assume that it pays off to use non-literal speech instead of conventional. We, however, will model the extra gain by a parameter $\epsilon > 0$. On the other hand we argued that unsuccessful communication is more ‘painful’ for the interlocutor (possibly both) who deviated from the convention. So if **S** uses message $m \neq \mathbf{cs}(t)$ (e.g. $m = \text{“wanker”}$) to communicate t (e.g. $t = \text{“Hey friend”}$), but **R** performs $a = \mathbf{cr}(\mathbf{cs}(t))$ (e.g. hitting **S** in the face), **S** and **R** did not successfully communicate. We hold that this is more painful for **S** than it is for **R**: **R** gets 0 as usual, **S** gets $-\epsilon'$, for some parameter value $\epsilon' > 0$. From our discussion above it follows that $\epsilon' > \epsilon$.

This brings us to the main definition of this paper. A *Super Conventional signaling game* $\mathbf{G}_{\langle \mathbf{cs}, \mathbf{cr} \rangle}$ is a standard signaling game \mathbf{G} equipped with a convention

$$\langle \{\mathbf{S}, \mathbf{R}\}, \{\mathbf{S}, \mathbf{R}\}, \{U_{\mathbf{S}}, U_{\mathbf{R}}\}, \langle \mathbf{cs}, \mathbf{cr} \rangle \rangle,$$

where $U_{\mathbf{S}}$ and $U_{\mathbf{R}}$ are equal to (1.2) and (1.3), except for $u_{\mathbf{S}}$ and $u_{\mathbf{R}}$ as defined in (1.1), respectively. Instead of (1.1) we take

$$u_{\mathbf{S}}^{\mathbf{cs}}(f, t, \mathbf{s}, \mathbf{r}) = \begin{cases} 1 + \epsilon, & \text{if } f(t) = \mathbf{r}(\mathbf{s}(t)) \text{ and } \mathbf{s}(t) \neq \mathbf{cs}(t); \\ 1, & \text{if } f(t) = \mathbf{r}(\mathbf{s}(t)) \text{ and } \mathbf{s}(t) = \mathbf{cs}(t); \\ 0, & \text{if } f(t) \neq \mathbf{r}(\mathbf{s}(t)) \text{ and } \mathbf{s}(t) = \mathbf{cs}(t); \\ -\epsilon', & \text{if } f(t) \neq \mathbf{r}(\mathbf{s}(t)) \text{ and } \mathbf{s}(t) \neq \mathbf{cs}(t); \end{cases} \quad (1.4)$$

and

$$u_{\mathbf{R}}^{\text{cr}}(f, t, \mathbf{s}, \mathbf{r}) = \begin{cases} 1 + \epsilon, & \text{if } f(t) = \mathbf{r}(\mathbf{s}(t)) \text{ and } \mathbf{r}(\mathbf{s}(t)) \neq \mathbf{cr}(\mathbf{s}(t)); \\ 1, & \text{if } f(t) = \mathbf{r}(\mathbf{s}(t)) \text{ and } \mathbf{r}(\mathbf{s}(t)) = \mathbf{cr}(\mathbf{s}(t)); \\ 0, & \text{if } f(t) \neq \mathbf{r}(\mathbf{s}(t)) \text{ and } \mathbf{r}(\mathbf{s}(t)) = \mathbf{cr}(\mathbf{s}(t)); \\ -\epsilon', & \text{if } f(t) \neq \mathbf{r}(\mathbf{s}(t)) \text{ and } \mathbf{r}(\mathbf{s}(t)) \neq \mathbf{cr}(\mathbf{s}(t)). \end{cases} \quad (1.5)$$

In the remainder of this section we will prove that every signaling system is a Nash equilibrium; furthermore we give characterizations of pay-off dominance and risk dominance in SC signaling games with respect to signaling systems (and not Nash equilibria) to trigger Rules 1 and 2.

Fact 1 *Let T , M and A be sets such that $|T| = |M| = |A|$. Let $\mathbf{G}_{\langle \text{cs}, \text{cr} \rangle}$ be a SC signaling game. If $\langle \mathbf{s}, \mathbf{r} \rangle$ is a signaling system in $\mathbf{G}_{\langle \text{cs}, \text{cr} \rangle}$, then $\langle \mathbf{s}, \mathbf{r} \rangle$ is a Nash equilibrium in $\mathbf{G}_{\langle \text{cs}, \text{cr} \rangle}$.*

Proof The crux of the proof is the fact that both \mathbf{s} and \mathbf{r} are injective functions, assuming that $\langle \mathbf{s}, \mathbf{r} \rangle$ is a signaling system. Any unilateral deviation will result in a function that communicates at least one $t \in T$ incorrectly. Therefore, \mathbf{S} nor \mathbf{R} can gain from unilateral deviation. \square

The converse does not hold. But we point out that every Nash equilibrium $\langle \mathbf{s}, \mathbf{r} \rangle$, that is not a signaling system, can be *extended* to a signaling system in the following sense: Let $S \subset T$ be the set of types that is successfully communicated through $\langle \mathbf{s}, \mathbf{r} \rangle$: $S = \{t \in T \mid f(t) = \mathbf{r}(\mathbf{s}(t))\}$. Then, there exists a signaling system $\langle \mathbf{s}^\#, \mathbf{r}^\# \rangle$ such that $\mathbf{s}^\#(t) = \mathbf{s}(t)$, for every $t \in S$. Furthermore, we claim that every signaling system $\langle \mathbf{s}^\#, \mathbf{r}^\# \rangle$ thus obtained, pay-off dominates $\langle \mathbf{s}, \mathbf{r} \rangle$. The signaling systems characterized in Fact 2 (below) count as a special case of this claim.

As we saw in the proof of Fact 1, there exist equilibria that communicate some types literally and some non-literally. A signaling system $\langle \mathbf{s}, \mathbf{r} \rangle$ divides T in two disjoint subsets

$$L_{\langle \mathbf{s}, \mathbf{r} \rangle} = \{t \in T \mid \mathbf{s}(t) = \mathbf{cs}(t) \text{ and } f(t) = \mathbf{r}(\mathbf{s}(t))\}$$

and

$$M_{\langle \mathbf{s}, \mathbf{r} \rangle} = \{t \in T \mid \mathbf{s}(t) \neq \mathbf{cs}(t) \text{ and } f(t) = \mathbf{r}(\mathbf{s}(t))\}.$$

That is, $L_{\langle \mathbf{s}, \mathbf{r} \rangle}$ is the set of types that are communicated literally and $M_{\langle \mathbf{s}, \mathbf{r} \rangle}$ contains all other types — the ones communicated non-literally. We call the number of types that are communicated literally the *rank* of an equilibrium, relative to the game $\mathbf{G}_{\langle \text{cs}, \text{cr} \rangle}$ of course. Formally, the rank of a signaling system $\langle \mathbf{s}, \mathbf{r} \rangle$ equals $|L_{\langle \mathbf{s}, \mathbf{r} \rangle}|$. It is an easy exercise to prove that for all $r \in \{0, \dots, |T|\}$ there exists a signaling system with rank r . The pay-offs of a signaling system can be expressed in terms of these sets $L_{\langle \mathbf{s}, \mathbf{r} \rangle}$ and $M_{\langle \mathbf{s}, \mathbf{r} \rangle}$. In a signaling system $\langle \mathbf{s}, \mathbf{r} \rangle$ namely, \mathbf{S} gains 1 or $1 + \epsilon$ per $t \in T$ depending on whether $t \in L_{\langle \mathbf{s}, \mathbf{r} \rangle}$ or $t \in M_{\langle \mathbf{s}, \mathbf{r} \rangle}$. Therefore,

$$U_S(\mathbf{s}, \mathbf{r}) = \left(\sum_{t \in L_{\langle \mathbf{s}, \mathbf{r} \rangle}} \mathbf{P}(t) \right) + \left(\sum_{t \in M_{\langle \mathbf{s}, \mathbf{r} \rangle}} (1 + \epsilon) \mathbf{P}(t) \right).$$

Since $\langle \mathbf{s}, \mathbf{r} \rangle$ is a signaling system, \mathbf{s} is injective; hence the conditional probability that t is selected, provided that message $\mathbf{s}(t)$ is sent, is 1. Consequently, $U_S(\mathbf{s}, \mathbf{r}) = U_R(\mathbf{s}, \mathbf{r})$ for all signaling systems $\langle \mathbf{s}, \mathbf{r} \rangle$.

Rule 1 and 2 mention pay-off and risk dominant Nash equilibria. To see what strategy-pairs have this property in SC signaling games we give characterizations of both notions in Fact 2 and 3. Fact 3 only characterizes risk dominant signaling systems; from our above discussion concerning extensions of equilibria we hope to have made clear that in fact it is the signaling systems that are most interesting anyhow.

Fact 2 *Let $\langle \mathbf{s}, \mathbf{r} \rangle$ be a Nash equilibrium in $\mathbf{G}_{\langle \mathbf{cs}, \mathbf{cr} \rangle}$. Then $\langle \mathbf{s}, \mathbf{r} \rangle$ is a pay-off dominant Nash equilibrium in $\mathbf{G}_{\langle \mathbf{cs}, \mathbf{cr} \rangle}$ iff $\langle \mathbf{s}, \mathbf{r} \rangle$ is a signaling system of rank 0.*

Proof Realize that the maximum pay-off is $1 + \epsilon$. □

Fact 3 *Let $\langle \mathbf{s}, \mathbf{r} \rangle$ be a signaling system in $\mathbf{G}_{\langle \mathbf{cs}, \mathbf{cr} \rangle}$. If $\epsilon' > \epsilon$, then $\langle \mathbf{s}, \mathbf{r} \rangle$ risk dominates all other signaling systems in $\mathbf{G}_{\langle \mathbf{cs}, \mathbf{cr} \rangle}$ iff $\mathbf{s} = \mathbf{cs}$ and $\mathbf{r} = \mathbf{cr}$.*

Proof (\Leftarrow) We observe that $L_{\langle \mathbf{cs}, \mathbf{cr} \rangle} = T$. We will prove that for any other signaling system $\langle \mathbf{s}, \mathbf{r} \rangle$ in $\mathbf{G}_{\langle \mathbf{cs}, \mathbf{cr} \rangle}$, we have that

$$\begin{aligned} A &= (U_S(\mathbf{cs}, \mathbf{cr}) - U_S(\mathbf{s}, \mathbf{cr}))(U_R(\mathbf{cs}, \mathbf{cr}) - U_R(\mathbf{cs}, \mathbf{r})) \\ &\quad \geq \\ &= (U_S(\mathbf{s}, \mathbf{r}) - U_S(\mathbf{cs}, \mathbf{r}))(U_R(\mathbf{s}, \mathbf{r}) - U_R(\mathbf{s}, \mathbf{cr})) = B. \end{aligned}$$

If a conventional strategy, say \mathbf{cs} , is played against a non-literal one, say \mathbf{r} , communication fails in case Nature selected a type t that is non-literally interpreted by \mathbf{r} ; that is, $t \in M_{\langle \mathbf{s}, \mathbf{r} \rangle}$. We denote

$$\left(\sum_{t \in L_{\langle \mathbf{s}, \mathbf{r} \rangle}} \mathbf{P}(t) \right) \text{ and } \left(\sum_{t \in M_{\langle \mathbf{s}, \mathbf{r} \rangle}} \mathbf{P}(t) \right)$$

by p and $1 - p$, respectively. The respective pay-offs can be obtained from (1), (2), (4) and (5). Filling those in yields: $A = (1 - (p - \epsilon'(1 - p)))^2$ and $B = ((1 + \epsilon)(1 - p))^2$. And, as required, by rewriting we learn that $A > B$, if $\epsilon' > \epsilon$.

(\Rightarrow) Suppose $\epsilon' > \epsilon$ and $\langle \mathbf{s}, \mathbf{r} \rangle$ is not risk dominant. Then this signaling system is not equal to the conventional signaling system, since this one was risk dominant. □

If we conceive of a SC signaling game as a real game, played by real people in a lab, we can apply Rule 1 and 2 from Section 3 to predict the players' behavior. That is, following

Rule 1, people by default coordinate on a risk dominant equilibrium. Applying Fact 3 yields that thus people by default communicate by means of the conventional signaling system. This matches Rule i. Now, if sympathy increases between people (and their common ground increases) a pay-off dominant, risk dominated equilibrium is more likely to be realized. From Fact 2 we learn that realizing a pay-off dominant equilibrium is equal to realizing a signaling system that communicates non-literally. Again this matches Rule ii.

5 Conclusion

In this papers we have introduced a model for non-literal language usage, that builds on Lewisian signaling games. Furthermore, it is consistent with Searle's conception of non-literal language. On the other hand, the model meets the rawest desiderata set by experimental literature on coordination games and language use.

For future research we suggest to embed our game-theoretical model in a broader pragmatical context. A more detailed study of the role of sympathy within the model as well as a more detailed account of different kinds of non-literal speech acts would be interesting.

Bibliography

- Harsanyi, J. and R. Selten (1988). *A General Theory of Equilibrium Selection in Games*. Cambridge: MIT Press.
- Lewis, D. (1969). *Convention*. Cambridge, Massachusetts: Harvard University Press.
- Osborne, M. and A. Rubinstein (1994). *A Course in Game Theory*. Cambridge: The MIT Press.
- Sally, D. Can I say, ‘Bobobo’ and mean, ‘There’s no such thing as cheap talk’? forthcoming in *Journal of Economic Behavior and Organization*.
- Sally, D. (2003). Risky speech: behavioral game theory and pragmatics. *Journal of Pragmatics* 35, 1223–1245.
- Searle, J. (1979). *Metaphor*, Chapter 6, pp. 92–123. Cambridge: Cambridge University Press.
- van Rooij, R. (2004). Signaling games select horn strategies. *Linguistics and Philosophy* 27, 493–527.