The Role of Hearers' Beliefs in the Interpretation of Logical Connectives

Logical connectives like *'or'* and *'if-then'* have been at the center of research on conversational implicatures since Grice. Under the standard account, a sentence like *'The strawberry is next to an apple or a lemon'* entails that the strawberry is next to an apple or a lemon or both, and implicates that it is not next to both. It is widely assumed that the successful communication of conversational implicatures depends on (i) alternative expressions the speaker could have used but didn't, and (ii) mutual beliefs about the goals and rationality of the interlocutors. This makes implicature a prime object for the application of game-theoretical methods, a project pursued by Parikh (1992, 2001) and others. In addition, psycholinguists have begun to study the processing of implicatures (Noveck *etal.*, 2002; Noveck and Chierchia,2005; Storto and Tanenhaus, 2005; among others). However, this work has not fully exploited what we consider a major strength of the game-theoretical approach, namely the fact that it delivers precise predictions about speakers' and listeners' behavior not only in standard communicative situations, but also under systematic manipulations of certain contextual factors that are deemed relevant, such as hearers' beliefs about speakers' interests. This paper presents a theoretical investigation and the results of an experimental pilot study aimed at filling this gap.

Theoretical background. This study is concerned with connectives like *'or'* and *'if'*. (In this abstract, we only discuss the former.) A Parkikh-style game-theoretical account of the implicatures of *'or'*-sentences assumes that interlocutors share three types of information: (i) the utilities of successful and unsuccessful communication of a given proposition; (ii) the costs associated with the use of particular sentences; and (iii) the probabilities of various possible speaker intentions. We extend this model to show how interlocutors can use such knowledge to choose a strategy for deciding between the interpretations of *'or'*-sentences in different situations, such as the case that the speaker knows enough to use an unambiguous alternative but who will not or cannot do so, and the case that the speaker must speak truthfully but stands to gain from miscommunication.

We call the inclusive and exclusive interpretation 'ι' and 'ϵ', respectively. Following Grice, the task is not semantic disambiguation but to decide whether to strengthen the semantic meaning with the exclusivity implicature. Typically, a speaker will use the form *'P or Q'* if he has incomplete evidence (e.g., if he knows that one of $p\overline{q}$, $\overline{p}q$, pq holds, but does not know which one). Suppose, in contrast, that he knows which of the three holds, and intends to communicate what he knows. We represent these situations as $s_{p\overline{q}}$, $s_{\overline{p}q}$ and s_{pq}. It can be shown that if there are unambiguous alternative forms $\mu_{p\overline{q}}$, $\mu_{\overline{p}q}$ and μ_{pq}, an uninformative expression like *'P or Q'* will not be used in any of the situations s_α unless it is less costly than the corresponding μ_α. Thus, a speaker must have some reason for being vague if he knows the whole truth. In the application of the model behind our experiment design, we stipulate that (i) the speaker knows which of the three states of affairs holds; but (ii) is barred from using unambiguous forms; (iii) that he nevertheless speaks the truth; and (iv) that these facts are mutually known. This sets up a non-standard situation with regard to the Gricean rationale for the exclusivity implicature, but one for which the model nevertheless makes precise predictions.

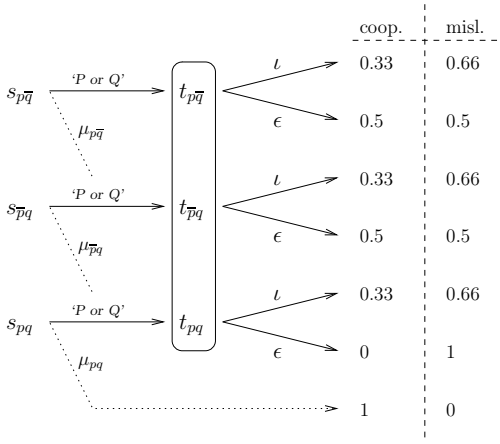

Figure 1: Expected payoff for cooperative and misleading speakers using *'P or Q'* with listeners choosing inclusive (ι) and exclusive (ϵ) readings

Supposing that the speaker has reasons for using *'P or Q'* despite knowing more, and that he speaks truthfully, we represent the situation as in Figure 1. On the left are the propositions that, for all the hearer knows, the speaker may be trying to communicate. The first set of arrows corresponds to the speaker's uttering the sentence. Thus $t_{p\overline{q}}$, $t_{\overline{p}q}$ and t_{pq} are just like their respective ancestors, except that the sentence has been uttered. The listener, who could not distinguish between the s-states, cannot distinguish between the t-states either. (She can rule out $t_{\overline{p}\,\overline{q}}$, which we omit here.) The next transition represents the the listener's choices and the corresponding payoffs.

Unlike in Parikh's model, the speaker's intentions ($p\overline{q}$, $\overline{p}q$, pq) and the listener's interpretations (ι, ϵ) are given different representations here. The latter correspond to different ways in which she may narrow down her choice set from among the possible intentions. These are not the only subsets she could have constructed (singletons are not ruled out), but we assume that these readings are saliently associated with *'or'*. However, the ultimate move in the game is for the listener to guess which of the three propositions ($p\overline{q}$, $\overline{p}q$, pq) the speaker in fact intended. Thus there are two steps to the interpretation: one consists in narrowing down the choice set, the other in choosing a specific proposition.

Our payoff structure also differs from Parikh's. For simplicity, we assume that miscommunication results in a forgone positive payoff, rather than incurring a cost. We also assume that the cost of uttering *'P or Q'* is neglible, thus an unsuccessful use has a net payoff of zero. These assumptions afford us some formal simplifications while preserving two crucial aspects of the model: (i) successful communication using $\mu_{p\overline{q}}$, $\mu_{\overline{p}q}$ or μ_{pq} is more costly than successful communication using *'P or Q'*; and (ii) successful communication using *'P or Q'* has a higher payoff than unsuccessful communication using *'P or Q'*.

We assign a "unit" payoff to successful communication: If the hearer is in $t_{p\overline{q}}$ and chooses $p\overline{q}$, both receive +1. Figure 1, however, reflects the assumption that the hearer first restricts her choice set to either ι or ϵ, and then chooses from that set at random. Her choice between ι and ϵ therefore determines the probability that she will correctly identify the speaker's intention, and thus her expected payoff. If she is in $t_{p\overline{q}}$, $t_{\overline{p}q}$ or t_{pq} and chooses ι, she will be choosing at chance between $s_{p\overline{q}}$, $s_{\overline{p}q}$ and s_{pq}, so her expected payoff is +0.33. If she is in $t_{p\overline{q}}$ or $t_{\overline{p}q}$ and chooses ϵ, her expected payoff is +0.5. If she chooses ϵ in t_{pq}, however, the payoff is always 0, since her choice set excludes the speaker's actual intention.

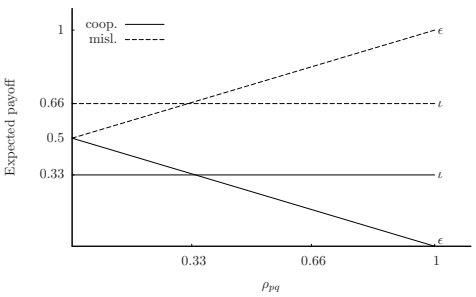

Figure 2: Expected payoff using *'p or q'* inclusively (ι) or exclusively (ϵ) as a function of the probability of *pq*

Finally, $\rho_{p\overline{q}}$, $\rho_{\overline{p}q}$, and ρ_{pq} are the respective probabilities that each of the intentions is the speaker's actual one. They sum to 1, since the three possibilities are mutually exclusive and jointly exhaustive ways for a fully informed speaker to use *'P or Q'* truthfully. Notice that they reflect only what is publicly known by the hearer, since the speaker would always assign a probability of 1 to his actual intention. In Parikh's proposal, these probabilities must be independent of the utterance, though we make no such assumption here, since we are artificially restricting the speaker to using *'P or Q'*.

Under these assumptions, we can evaluate strategies for interpreting *'or'*. (We do not consider mixed strategies.) First, the expected utility of each strategy (i.e., interpreting *'P or Q'* as ι versus ϵ) varies with the probability distribution:

$$EU(\iota) = 0.33 \times \rho_{p\overline{q}} + 0.33 \times \rho_{\overline{p}q} + 0.33 \times \rho_{pq} = 0.33 \tag{1}$$

$$EU(\epsilon) = 0.5 \times \rho_{p\overline{q}} + 0.5 \times \rho_{\overline{p}q} + 0 \times \rho_{pq} = 0.5 \times (\rho_{p\overline{q}} + \rho_{\overline{p}q}) = 0.5 \times (1 - \rho_{pq}) \tag{2}$$

While $EU(\iota)$ does not depend on the probabilities, $EU(\epsilon)$ depends entirely on ρ_{pq}. Figure 2 (the solid lines) graphs this dependence. Thus the preferred strategy for interpreting *'P or Q'* depends on ρ_{pq}. If $\rho_{pq} > 0.33$, then $EU(\iota) > EU(\epsilon)$. If $\rho_{pq} < 0.33$, $EU(\epsilon) > EU(\iota)$. If $\rho_{pq} = 0.33$, neither strategy wins.

The picture changes if it is mutually known that the speaker, while being truthful, has an interest in *misleading* the hearer. In that case, we assign the same high cost to $\mu_{p\overline{q}}$, $\mu_{\overline{p}q}$ and μ_{pq}, but now the speaker receives a positive payoff only if the hearer chooses incorrectly. This structure, called "misleading" in Figures 1 and 2, gives rise to the following:

$$EU(\iota) = 0.66 \times \rho_{p\overline{q}} + 0.66 \times \rho_{\overline{p}q} + 0.66 \times \rho_{pq} = 0.66 \tag{3}$$

$$EU(\epsilon) = 0.5 \times \rho_{p\overline{q}} + 0.5 \times \rho_{\overline{p}q} + 1 \times \rho_{pq} = 0.5 \times (\rho_{p\overline{q}} + \rho_{\overline{p}q}) + \rho_{pq} = 0.5 \times \rho_{pq} + 0.5 \tag{4}$$

The pattern of success on the part of the hearer does not differ in this case, but since the speaker now benefits whenever the hearer makes the wrong choice, his expected payoff under ι is $1 - 0.33 = +0.66$. Under ϵ, his expected payoff is still +0.5 in $t_{p\overline{q}}$ and $t_{\overline{p}q}$, but in t_{pq}, the expected payoff is +1. Now a different pattern of preference emerges, shown in Figure 2. If $\rho_{pq} < 0.33$, the speaker prefers ι. If $\rho_{pq} > 0.33$, he prefers ϵ. In sum, opposite patterns are predicted depending on whether the speaker wants his intention to be revealed or not.

Experiment. Unlike much earlier experimental research on implicatures, the question our pilot study addresses is directly related to their game-theoretical rationale: Will subjects' behavior change as predicted under systematic manipulations of such crucial variables as the payoff structure?

Method. Fifteen undergraduate students enrolled at Northwestern University participated in this experiment for course credit. They were asked to aid two fictional characters in playing a series of games. For each game, they were shown a grid with pictures of various objects (see Figure 3) on a computer screen. In each game, the fictional characters had hidden a prize behind one of the objects (the "target"). The task was to find the target based on a verbal clue from one of the players. The rules stipulated that the clues were to be truthful but not completely informative. There were three types of clues, dubbed *'or'*, *'not-and'*, and *'if'*; examples are in (5). (Notice that (5c) corresponds to *'if Q then P'*. This form was chosen so as to maximize the parallelism between the stimuli.)

(5) a. The prize is behind a strawberry that is next to an apple or a lemon.
b. The prize is behind a strawberry that is not next to an apple and a lemon.
c. The prize is behind a strawberry that is next to an apple if it is next to a lemon.

Each grid contained four potential targets, exhibiting all logical possibilities. Participants were told that "next to" meant horizontally or vertically adjacent, not diagonally, and that one player (the "winner") would get the prize if they chose correctly; otherwise, the other player (the "loser") would get the prize. Each clue was presented visually as coming from either the winner or the loser.

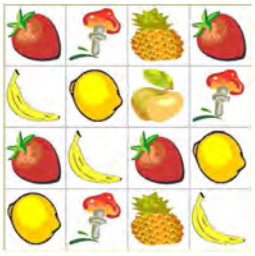

Figure 3: Sample grid

Results. More than half of the participants chose pq for both *'or'* and *'if-then'*, and more than half chose $\overline{p}\,\overline{q}$ for *'not-and'*. All of these patterns were significant (1-way ANOVA, *'or'*: $F(2,28) = 5.15, p < .05$; *'if-then'*: $F(2,28) = 27.97, p < .001$; *'and-not'*: $F(2,28) = 122.03, p < .001$). Furthermore, a 2-way ANOVA revealed a significant interaction between type of utterance, speaker role, and pattern of response ($F(8,112) = 2.081, p < .05$). The different patterns are shown in Figure 4. In addition, there was a significant 2-way interaction of the utterance type and the pattern of responses ($F(8,112) = 33.31, p < .001$). The 2-way interaction between speaker role and response pattern was not significant ($F(4,56) < 1, n.s.$). Lastly, the response pattern also exhibited a significant main effect ($F(4,56) = 38.01, p < .001$). In order to further investigate the 3-way interaction, three planned 2-way ANOVAs were conducted, one for each utterance type, the independent variables being speaker role and responses pattern. The interaction between response patterns and speaker role was significant for *'if-then'*-type utterances ($F(4,56) = 3.01, p < .05$). For the other forms, it was present as a trend but not significant.

Discussion. One result that stands out is that listeners overwhelmingly chose pq in the case of both *'or'* and *'if'*, suggesting that they interpreted the connectives inclusively. According to our model, this is predicted to occur only when the probability of the pragmatically "excluded" proposition (e.g., s_{pq} for *'or'*) is mutually believed to be high. Moreover, the fact that subjects preferred one of the three possibilities over the other two by a wide margin suggests that they had evidence for a relatively high probability of this particular choice. But recall that our experimental setup did not introduce such a bias: All three objects were equally likely to be the one with the prize. We may then surmise that subjects' biases arise from the fact that the speaker used the sentence.

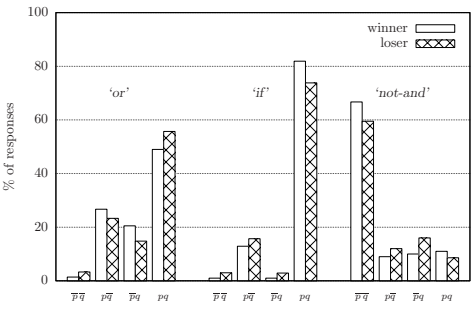

Figure 4: Percent of responses for different utterance types and speaker roles

Why should the speaker's use of the sentence have this effect? Notice that this cannot be explained in Gricean terms on logical grounds alone. Grice would predict *'P if Q'* in a setting like ours to lose its implicature and become equivalent to the material conditional *'P or not-Q'*. But then by analogy with *'P or Q'*, we would expect subjects to prefer $p\overline{q}$, contrary to our results.

An alternative explanation would be that listeners are simply drawn towards the object that is next to both of the ones mentioned in the sentence. (For the fact that this is not the case with *'not-and'*, see below.) What remains to be seen is whether this is due to an intervention of psychological processes (e.g., priming due to their being mentioned) or part of a conventional strategy speakers and listeners use to deal with situations like ours. Further work is required to address this question.

Compared to these rather pronounced patterns, the effect of the role of the speaker ("winner" *vs.* "loser") was rather small, but here, too, we see interesting trends. Even with our small number of subjects, the effect was significant in the case of *'if'* and present numerically for *'or'*; however, the two are reversed. Subjects tended to give *'or'* an exclusive reading when the speaker stood to gain from a successful communication, but did the opposite otherwise. This is compatible with an interpretation in which listeners expect the "loser" to resort to a non-conventional use of the utterance in order to deceive them. This generalizes to the case of *'if'* under the above assumption that the pq case is the "conventional" one in this case.

However, the interpretation of *'not-and'* sentences is not as clear. In particular, the proportion of inappropriate responses (i.e. pq objects) is on par with the proportion of the non-modal appropriate responses ($\overline{p}q$, $p\overline{q}$). This suggests that participants had a difficult time interpreting these utterances, perhaps due to the interference of psychological processes. Specifically, the high proportion of errors suggested that in some cases participants may have missed the negation operator, possibly due to the complexity of the utterance.