Interpretation of Optimal Signals

Michael Franke[*]

Institute for Logic, Language and Computation
Universiteit van Amsterdam
Nieuwe Doelenstraat 15
1012 CP Amsterdam, The Netherlands
m.franke@uva.nl

Abstract

According to the optimal assertions approach of Benz and van Rooij (2007), conversational implicatures can be calculated based on the assumption that a given signal was optimal, i.e. that it was the sender's best choice if she assumes, purely hypothetically, a particular naive receiver interpretation behavior. This paper embeds the optimal assertions approach in a general signaling game setting and derives the notion of an optimal signal via a serious of iterated best responses (c.f. Jäger, 2007). Subsequently, we will compare three different ways of interpreting such optimal signals. It turns out that under a natural assumption of expressibility (i) the optimal assertions approach, (ii) iterated best response and (iii) strong bidirectional optimality theory (Blutner, 1998, 2000) all prove equivalent. We then proceed to show that, if we take the iterated best response sequence one step further, we can account for M-implicatures (Horn's division of pragmatic labor) standardly in terms of signaling games.

Often we express more with the use of our words than what those words mean literally. For example, if you were to say that this observation is not particularly new, I would clearly get the hint and understand that you meant to say that it is more than just not particularly new, indeed a working standard in linguistic pragmatics. Such *conversational implicatures* were first studied by Grice (1989) and still concern the community in various ways. In particular, recent years saw an increasing interest in game-theoretical models of conversational implicature calculation, and this study belongs to this line of research. It provides a formal comparison of selected previous approaches

[*] Thanks to Anton Benz, Reinhard Blutner, Tikitu de Jager, Gerhard Jäger, Robert van Rooij and an anonymous referee for very helpful comments and discussions.

which extends to a uniform synchronic account of different kinds of conversational implicatures.

The paper is organized as follows. Section 1 briefly reviews the classification of conversational implicatures into I-, Q- and M-implicatures. Section 2 introduces a game-theoretical model of implicature calculation: a signaling game with exogenously meaningful signals. We will see in section 2.3 that the standard solution concept for signaling games is not strong enough to account for the empirical observations. The *optimal assertions approach* of Benz and van Rooij (2007), which is introduced in section 3.1, is an attempt to solve this problem. According to the optimal assertions approach, conversational implicatures can be calculated based on the assumption that a given signal was *optimal*. Sections 3.2 and 3.3 then compare three ways of interpreting such optimal signals: (i) the pragmatic interpretation rule of Benz and van Rooij (2007), (ii) iterated best response and (iii) strong bidirectional optimality theory (Blutner, 1998, 2000). It turns out that if we assume a sufficiently expressible stock of possible signals, all three approaches prove equivalent. However, it also turns out that M-implicatures (Horn's division of pragmatic labor) cannot be accounted for based solely on the assumption that the received form was optimal. We will conclude that some aid from the refinement literature, in particular Cho and Kreps' (1987) intuitive criterion, is necessary and sufficient to account uniformly for all I-, Q- and M-implicatures.

1 Kinds of Conversational Implicatures

Neo-Gricean pragmatics (Atlas and Levinson, 1981; Horn, 1984) distinguishes I-implicatures (1) and Q-implicatures (2).

(1) John has a very efficient secretary.
$\rightsquigarrow$ John has a very efficient *female* secretary.

(2) John invited some of his friends.
$\rightsquigarrow$ John did not invite all of his friends.

I-implicatures like (1) are inferences to a stereotype: the sentence is associated with the most likely situation consistent with its semantic meaning. Q-implicatures like (2), also called scalar implicatures, are a strengthening of the literal meaning due to the presence of more informative alternatives that were not used: since the speaker only said that some of John's friends were invited, we infer that the compatible stronger claim that all of John's friends were invited — a claim that we may assume relevant if true — does not hold, for otherwise the speaker would have said so — as she is assumed cooperative and informed.

A third kind of implicature, called M-implicature by Levinson (2000), is given in (3).

(3) The corners of Sue's lips turned slightly upwards.
$\rightsquigarrow$ Sue didn't smile genuinely, but faked a smile.

In (3) we naturally infer that something about the way Sue smiled was abnormal, non-stereotypical or non-standard, because the speaker used a long and complicated form where she could have used the simple expression (4).

(4) Sue smiled.

M-implicatures were also discussed by Horn (1984) and have been addressed as *Horn's division of pragmatic labor* thereafter. It has become customary to assume that both sentences (3) and (4) are semantically equivalent, but, when put to use, the longer form (3) gets to be associated with the non-stereotypical situation, while the short form (4) gets to be associated with the stereotypical situation.

2 Implicatures via Signaling Games

2.1 Interpretation Frames

A fairly manageable set of contextual parameters plays a role in the neo-Gricean classification of implicatures: we distinguish various meanings that are more or less stereotypical and we compare different forms with respect to their semantic meaning and complexity. We can then capture any such configuration of contextual parameters that are relevant for the computation of implicatures in an *interpretation frame.*

Definition 2.1 (Interpretation Frame)**.** An interpretation frame is a tuple

$$\mathcal{F} \stackrel{\text{def}}{=} \langle W, P, F, c, [\![\cdot]\!] \rangle$$

where W is a finite set of worlds or situations, P is a probability distribution over W with the usual properties,[1] F is a set of forms or signals which the sender may send, $c : F \rightarrow \mathbb{R}$ is a cost function and $[\![\cdot]\!] : F \rightarrow \mathscr{P}(W)$ is a semantic denotation function mapping forms to subsets of W.

[1] $P(w) \in [0,1]$, for all $w \in W$; $P(A) = \sum_{w \in A} P(w)$, for all $A \subseteq W$; $P(W) = 1$.

We assume for convenience that $P(w) \neq 0$ for all worlds $w \in W$. We would also like to rule out certain rather pathological situations where there are worlds which simply cannot be expressed by any conventional signal:

Assumption 2.2 (Semantic Expressibility)**.** We only consider interpretation frames in which all worlds are semantically expressible: for all worlds w there has to be a form f such that $w \in [\![f]\!]$.

The kinds of implicatures described in the previous section correspond to abstract interpretation frames as follows:

- The *I-frame* is an interpretation frame $\mathcal{F}_I = \langle W, P, F, c, [\![\cdot]\!]\rangle$ where $W = \{w, v\}$, $P(w) > P(v) \neq 0$, $F = \{f, g, h\}$, $c(f) < c(g), c(h)$ and $[\![f]\!] = W$, $[\![g]\!] = \{v\}$ and $[\![h]\!] = \{w\}$. The observed *I-implicature play* is to interpret f as w and to send f in w only.
- The *Q-frame* is an interpretation frame $\mathcal{F}_Q = \langle W, P, F, c, [\![\cdot]\!]\rangle$ where $W = \{w, v\}$, $P(w) \geq P(v) \neq 0$, $F = \{f, g\}$, $c(f) = c(g)$ and $[\![f]\!] = W$, $[\![g]\!] = \{v\}$. The observed *Q-implicature play* is to interpret f as w and to send f in w only.
- The *M-frame* is an interpretation frame $\mathcal{F}_M = \langle W, P, F, c, [\![\cdot]\!]\rangle$ where $W = \{w, v\}$, $P(w) > P(v) \neq 0$, $F = \{f, g\}$, $c(f) < c(g)$ and $[\![f]\!] = [\![g]\!] = W$. The observed *M-implicature play* is to interpret f as w and to send f in w only, as well as to interpret g as v and to send g in v only.

2.2 Interpretation Games

Interpretation frames capture the relevant aspects of the situation in which communication takes place. The communication itself can best be imagined as a signaling game: nature selects a world $w \in W$ — call it the actual world in a given play — with probability $P(w)$ and reveals it to the sender who in turn chooses a form $f \in F$. The receiver does not observe the actual world, but observes the signal f. He then chooses an action A. Sender and receiver receive a payoff based on w, f and A. In the present context, we are interested in *interpretation games*: signaling games in which signals have a conventional, compelling meaning that the receiver tries to interpret by choosing an interpretation action $\emptyset \neq A \subseteq W$.

Definition 2.3 (Interpretation Game)**.** An interpretation game is just an interpretation frame to which interpretation actions and utilities for sender and receiver are added, in other words a tuple

$$\mathcal{G} \stackrel{\text{def}}{=} \langle \mathcal{F}, Act, u_S, u_R \rangle$$

where $\mathcal{F} = \langle W, P, F, c, [\![\cdot]\!] \rangle$ is an interpretation frame, $Act \stackrel{\text{def}}{=} \mathscr{P}(W) \setminus \emptyset$ is a set of *interpretation actions* and $u_x : F \times Act \times W$ are utility functions of sender and receiver:[2]

$$\begin{aligned} u_R(f, A, w) &\stackrel{\text{def}}{=} \begin{cases} \frac{1}{|A|} & \text{if } w \in A \text{ and } w \in [\![f]\!] \\ 0 & \text{if } w \notin A \text{ and } w \in [\![f]\!] \\ -1 & \text{otherwise} \end{cases} \\ u_S(f, A, w) &\stackrel{\text{def}}{=} u_R(f, A, w) - c(f). \end{aligned}$$

As usual, we identify the receiver's probabilistic beliefs with the probability distribution $P(\cdot)$. Costs are assumed *nominal*: they are small enough to make a utility difference for the sender for any two different signals f and f' only in case $u_R(f, A, w) = u_R(f', A, w)$.

Definition 2.4 (Strategies)**.** A *sender strategy* is a function $\sigma : W \to \mathscr{P}(F) \setminus \emptyset$ that specifies a set $\sigma(w) \subseteq F$ of messages to be sent with equal probability when in world w. We call a sender strategy σ *truth-respecting* iff for all world w and f whenever $f \in \sigma(w)$ we have $w \in [\![f]\!]$. We define also $\sigma^{-1}(f) \stackrel{\text{def}}{=} \{w \in W \mid f \in \sigma(w)\}$. Finally, a *receiver strategy* is a function $\rho : F \to Act$ specifying an interpretation for each message.

Whether an action is preferable to another depends on what the other party is doing. If we fix a strategy for the other party we can define the expected utility of each action.

Definition 2.5 (Expected Utilities)**.** Since the sender knows the actual world w, his expected utility of sending the form $f \in F$ given that the receiver plays ρ is actually just his utility in w given f and the receiver's response $\rho(f)$:

$$\mathrm{EU}_S(f, \rho, w) \stackrel{\text{def}}{=} u_S(f, \rho(f), w).$$

[2] These utilities reflect the mutual desire to communicate which world is actual: the more the receiver narrows down a correct guess the better; miscommunication, on the other hand, is penalized so that if the chosen interpretation does not include the actual situation, the payoff is strictly smaller than when it does; a strong penalty is given for communication that deviates from the semantic meaning of messages to enforce the exogenous meaning of signals. (This last point is objectionable, but it is also not strictly necessary. I adopt it for ease of exposition since space is limited.)

Given that the sender plays σ, the receiver's expected utility of interpreting a form f for which $\sigma^{-1}(f) \neq \emptyset$ as $A \in Act$ is:[3]

$$\mathrm{EU}_R(A, \sigma, f) \overset{\text{def}}{=} \sum_{w \in W} P(w|\sigma^{-1}(f)) \times u_R(f, A, w)$$

For a truth-respecting sender strategy this simplifies to:

$$\mathrm{EU}_R(A, \sigma, f) = \frac{P(A|\sigma^{-1}(f))}{|A|}. \tag{2.1}$$

If the other party's strategy is given, rationality requires to maximize expected utility. A strategy X that maximizes expected utility in all its moves given the other party's strategy Y is called a *best response* to Y. For some sender strategies σ and forms f it may be the case that several actions maximize the receiver's expected utility, and that therefore there is no unique best response. Given 2.1, it is easy to see that all (non-empty) sets that contain only worlds which are maximally likely according to $P(\cdot|\sigma^{-1}(f))$ are equally good interpretations in expectation:[4]

$$\mathrm{Max}_{A \in Act} \mathrm{EU}_R(A, \sigma, f) = \mathscr{P}(\mathrm{Max}_{w \in W} P(w|\sigma^{-1}(f)))) \setminus \emptyset.$$

Assumption 2.6 (Preferred Interpretation)**.** We assume that the receiver selects as his best response to a truth-respecting σ and f the largest interpretation action $\mathrm{Max}_{w \in W} P(w|\sigma^{-1}(f))$. This is because the receiver should not discard any possible interpretation without reason; one should not gamble on proper understanding.[5]

The standard solution concept for rational play in a signaling game is a perfect Bayesian equilibrium: a pair of strategies that are best responses to one another.

Definition 2.7 (Perfect Bayesian Equilibrium)**.** A pair of strategies $\langle \sigma, \rho \rangle$ is a perfect Bayesian equilibrium iff

(i) for all $w \in W$: $\sigma(w) \in \mathrm{Max}_{f \in F} \mathrm{EU}_S(f, \rho, w)$

(ii) for all $f \in F$: $\rho(f) \in \mathrm{Max}_{A \in Act} \mathrm{EU}_R(A, \sigma, f)$.

[3] We will come back to the question how to interpret messages f in the light of sender strategies σ that never use f in sections 3.2 and 3.4. For the time being, assume that $\mathrm{EU}_R(A, \sigma, f) = 0$ is constant for all A if $\sigma^{-1}(f) = \emptyset$.

[4] We write $\mathrm{Max}_{x \in X} F(x) \overset{\text{def}}{=} \{x \in X \mid \neg\exists x' \in X : F(x) < F(x')\}$, for arbitrary set X and function $F : X \rightarrow \mathbb{R}$.

[5] This assumption replaces the tie-break rule of Benz and van Rooij (2007).

2.3 Pragmatics & the Problem of Equilibrium Selection

It is easy to verify that I-, Q- and M-implicature play are all perfect Bayesian Equilibria (PBEs) in the corresponding interpretation games, but not uniquely so. Indeed, the straight-forward signaling games approach to implicature computation faces a *problem of equilibrium selection*: why is it that particular PBEs are observed and not others?

A natural way of answering this question is to formulate refinements of the assumed solution concept. An interesting proposal along these lines is given by van Rooij (2007) who observes that the Q-implicature play can be singled out as the unique *neologism proof* PBE (Farrell, 1993) and that the M-implicature play can be singled out with the help of Cho and Kreps' *intuitive criterion* (Cho and Kreps, 1987). We will pick up this latter idea in section 3.4. Notice, however, that van Rooij's approach deviates from a standard signaling game analysis, because in order to arrive at the desired prediction for the M-frame, van Rooij considers a transition from an interpretation frame with just the cheaper message f, to which at a later stage the more costly message g is added. The question remains whether we cannot account for the observed implicature plays in more conservative terms?

3 Association-Optimal Signaling

A recent framework that seeks to give a positive answer to this question is Benz and van Rooij's (2007) *optimal assertions approach.* The basic idea is that the receiver may compute implicatures based on the assumption that the signal he received was an *optimal assertion.* An optimal assertion in turn is the best response to a naive, hypothetical interpretation of messages that takes into account only the semantic meaning of the message and the probabilities of worlds. Benz and van Rooij describe their set-up as a sequence of decision problems: on the hypothesis that the receiver interprets signals in a certain, naive way, the sender will choose signals that are optimal given this receiver strategy and the receiver can then interpret messages as optimal.

Another way of looking at this process is as a sequence of *iterated best responses* (c.f. Jäger, 2007). To point out the connection, I will spell out the details of the optimal assertions approach in terms of iterated best responses in section 3.1. I will then, in section 3.2, show that Benz and van Rooij's interpretation rule deviates slightly from the former iterated best response logic in general, but that for a natural subclass of interpretation frames — including I- and Q-frames— the two approaches fall together. In section 3.3, finally, I will connect both the optimal assertion and the iterated best response approach with strong bidirectional optimality theory.

3.1 Association Optimality

We start with the assumption that the sender says something true:

$$\sigma_0(w) = \{f \in F \mid w \in [\![f]\!]\}\,.$$

We also assume that, given that the sender says something true, the receiver will interpret messages as true; in other words, as the sender starts with a naive 'truth-only' strategy σ_0, the receiver maximizes his expected utility based on that strategy and plays (as σ_0 is truth-respecting):

$$\begin{array}{rcl} \rho_0(f) & = & \text{Max}_{w \in W} P(w|\sigma_0^{-1}(f)) \\ & = & \text{Max}_{w \in W} P(w|\,[\![f]\!]). \end{array}$$

We could think here of a spontaneous, first associative response to the message f: the most likely worlds in which f is true are chosen as the first interpretation strategy, because these are the worlds that spring to mind first when hearing f. We therefore call ρ_0 the receiver's *association response.*

The association response ρ_0 is of course a bad interpretation strategy. In fact, it is not a pragmatic interpretation strategy at all, for it leaves out all considerations about the interpretation game except $[\![\cdot]\!]$ and $P(\cdot)$: receipt of message f is treated as if it was the observation of the event $[\![f]\!]$. But still the association response ρ_0 *is* the rational response to the — admittedly non-pragmatic — sender strategy σ_0. The guiding conviction here is that pragmatic reasoning takes semantic meaning as a starting point: if I want to know what you meant by a given linguistic sign, I first feed into the interpretation machine the conventional meaning of that sign. Therefore, as σ_0 is a natural beginning, so is the association response ρ_0.[6]

But if this truly is the most reasonable beginning for pragmatic interpretation, the sender may anticipate the receiver's association response ρ_0 and choose a best response to it:

$$\begin{array}{rcl} \sigma_1(w) & = & \text{Max}_{f \in F} \text{EU}_S(f, \rho_0, w) \\ & = & \{f \in F \mid \neg\exists f' \in F \;:\; \text{EU}_S(f, \rho_0, w) < \text{EU}_S(f', \rho_0, w)\} \end{array}$$

[6] An anonymous reviewer asks for the difference between Jäger's (2007) evolutionary model, which also uses best response dynamics, and the present synchronic approach. One obvious difference is that the present model assumes that at each turn a best response is selected with probability 1. Another difference is the starting point: in Jäger's model it is the sender, while in the present model it is receiver who responds first to a strategy that is given by the semantic meaning of the signals.

Forms in σ_1 are optimal forms given the receiver's association response. We could therefore call them *association optimal*, or, for short, *optimal:* a form $f \in F$ is (association) optimal in a world w iff $f \in \sigma_1(w)$.

How should the receiver interpret an optimal signal? We'll next consider and compare three possible answers to this question.

3.2 Optimal Assertions and Iterated Best Response

Given semantic expressibility as stated in assumption 2.2, association optimality is equivalent to Benz and van Rooij's (2007) notion of an optimal assertion. Although the latter notion requires truth of a message for its optimality, it is easy to see that semantic expressibility and optimality entail truth.

Observation 3.1. Given semantic expressibility, σ_1 is truth-respecting.

Proof. Let some $f \in F$ be false in $w \in W$. From semantic expressibility there is a message $f' \in F$ which is true in w. But then $-1 = u_S(f, \rho_0(f), w) < 0 \leq u_S(f', \rho_0(f'), w)$, so that f is not association optimal in w. Q.E.D.

If the sender sends an association optimal signal, i.e. if the sender sticks to σ_1, the receiver can again interpret accordingly. Benz and van Rooij propose the following interpretation rule based on the assumption that the received signal was an *O*ptimal *A*ssertion: $\rho_1^{\text{OA}}(f) = \{w \in [\![f]\!] \mid f \text{ is optimal in } w\}$. Thich simplifies under observation 3.1 to

$$\rho_1^{\text{OA}}(f) = \sigma_1^{-1}(f). \tag{3.1}$$

Notice, however, that this may not be a well-defined receiver strategy in our present set-up, for it may be the case that $\sigma_1^{-1}(f) = \emptyset$, which is not a feasible interpretation action. The same problem also occurs for the best response to σ_1. It is clear what the best response to σ_1 is for messages that may be optimal somewhere: if $\sigma_1^{-1}(f) \neq \emptyset$, we have

$$\rho_1^{\text{BR}}(f) = \text{Max}_{w \in W} P(w | \sigma_1^{-1}(f)). \tag{3.2}$$

But how should a best response to σ_1 interpret messages that are never optimal? Since we defined (tentatively, in footnote 3) expected utilities as constant for all $A \in \mathit{Act}$ whenever $\sigma^{-1}(f) = \emptyset$, any $A \in \mathit{Act}$ is an equally good interpretation for a non-optimal f. For our present purpose —the comparison of frameworks— it is not important what to choose in this case, as long as we choose consistently. We therefore adopt the following assumption and reflect on it in section 3.4 where it plays a crucial role.

Assumption 3.2 (Uninterpretability Assumption)**.** We assume that the receiver resorts to the mere semantic meaning in case a message is uninterpretable: if $\sigma_1^{-1}(f) = \emptyset$, then $\rho_1^{\mathrm{OA}}(f) = \rho_1^{\mathrm{BR}}(f) = [\![f]\!]$.

With this we can show that $\rho_1^{\mathrm{BR}}(f)$ entails $\rho_1^{\mathrm{OA}}(f)$ for arbitrary f and interpretation frames. Moreover, ρ_1^{OA} also entails ρ_1^{BR}, if we assume *strong expressibility*:

Definition 3.3 (Strong Expressibility)**.** We say that an interpretation frame satisfies strong expressibility if each world is immediately associated with some message: for each world w there is a form f such that $w \in \rho_0(f)$.

Observation 3.4. Under strong expressibility, association optimality implies inclusion in the association response: if f is association optimal in w, then $w \in \rho_0(f)$.

Proof. Assume strong expressibility. If $w \notin \rho_0(f)$, there is a form f' for which $w \in \rho_0(f')$. But then $0 = u_S(f, \rho_0(f), w) < u_S(f', \rho_0(f'), w)$. So f is not association optimal in w. Q.E.D.

Proposition 3.5. For arbitrary interpretation frames it holds that $\rho_1^{\mathrm{BR}}(f) \subseteq \rho_1^{\mathrm{OA}}(f)$. For interpretation frames satisfying strong expressibility it holds that $\rho_1^{\mathrm{BR}}(f) = \rho_1^{\mathrm{OA}}(f)$.

Proof. We only have to look at the non-trivial case where $\sigma_1^{-1}(f) \neq \emptyset$. Let $w \in \rho_1^{\mathrm{BR}}(f)$. Since all worlds have non-zero probabilities we can conclude that $w \in \sigma_1^{-1}(f)$. Hence, $w \in \rho_1^{\mathrm{OA}}(f)$.

Let $w \in \rho_1^{\mathrm{OA}}(f)$ and assume strong expressibility. Then $w \in [\![f]\!]$ and $f \in \sigma_1(w)$. From observation 3.4 we then know that $w \in \rho_0(f)$. That means that there is no w' for which $P(w' | [\![f]\!]) > P(w | [\![f]\!])$. But since, by observation 3.1, we know that $\sigma_1^{-1}(f) \subseteq [\![f]\!]$, we also know that there is no w' for which $P(w' | \sigma_1^{-1}(f)) > P(w | \sigma_1^{-1}(f))$. Hence $w \in \rho_1^{\mathrm{BR}}(f)$. Q.E.D.

3.3 Strong Bidirectional Optimality Theory

A similar connection holds with strong Bi-OT (Blutner, 1998, 2000). At first sight, Bi-OT looks rather different from game-theoretic models, because in Bi-OT we compare form-meaning pairs $\langle f, w \rangle$ with respect to a preference order. The idea is that to express a given meaning w with a form f, the form-meaning pair $\langle f, w \rangle$ has to be strongly optimal. Likewise, a form f will be associated with meaning w if and only if $\langle f, w \rangle$ is strongly optimal.

Definition 3.6 (Strong Bidirectional Optimality)**.** A form-meaning pair $\langle f, w\rangle$ is strongly optimal iff it satisfies both the Q- and the I-principle, where:

(i) $\langle f, w\rangle$ satisfies the Q-principle iff $\neg\exists f' : \langle f', w\rangle > \langle f, w\rangle$

(ii) $\langle f, w\rangle$ satisfies the I-principle iff $\neg\exists w' : \langle f, w'\rangle > \langle f, w\rangle$

How should we define preference relations against the background of an interpretation game? Recall that the Q-principle is a sender economy principle, while the I-principle is a hearer economy principle. We have already seen that each interlocutor's best strategy choice depends on what the other party is doing. So, given σ_0 and ρ_0 as a natural starting point we might want to define preferences simply in terms of expected utility:

$$\begin{aligned} \langle f', w\rangle > \langle f, w\rangle \quad &\text{iff} \quad \mathrm{EU}_S(f', \rho_0, w) > \mathrm{EU}_S(f, \rho_0, w) \\ \langle f, w'\rangle > \langle f, w\rangle \quad &\text{iff} \quad \mathrm{EU}_R(\{w'\}, \sigma_0, f) > \mathrm{EU}_R(\{w\}, \sigma_0, f) \end{aligned}$$

This simplifies to:[7]

$$\begin{aligned} \langle f', w\rangle > \langle f, w\rangle \quad &\text{iff} \quad u_S(f', \rho_0(f'), w) > u_S(f, \rho_0(f), w) \\ \langle f, w'\rangle > \langle f, w\rangle \quad &\text{iff} \quad P(w' | \llbracket f\rrbracket) > P(w | \llbracket f\rrbracket). \end{aligned}$$

Observation 3.7. Interpretation based on optimal assertions $\rho_1^{\mathrm{OA}}(f)$ is strong Bi-OT's Q-principle: a form-meaning pair $\langle f, w\rangle$ satisfies the Q-principle iff $\sigma_1^{-1}(f) \neq \emptyset$ and $w \in \rho_1^{\mathrm{OA}}(f)$.

Proof. A form-meaning pair $\langle f, w\rangle$ satisfies the Q principle iff there is no f' such that $\mathrm{EU}_S(f', \rho_0, w) > \mathrm{EU}_S(f, \rho_0, w)$ iff f is association optimal in w iff $\sigma_1^{-1}(f) \neq \emptyset$ and $w \in \rho_1^{\mathrm{OA}}(f)$. Q.E.D.

Let's capture interpretation based on strong optimality in an interpretation operator for ease of comparison. If $\sigma_1^{-1}(f) = \emptyset$, the uninterpretability assumption holds, and we take $\rho_1^{\mathrm{OT}}(f) = \llbracket f\rrbracket$; otherwise: $\rho_1^{\mathrm{OT}}(f) = \{w \in W | \langle f, w\rangle \text{ is strongly optimal}\}$, which is equivalent to:

$$\rho_1^{\mathrm{OT}}(f) = \{w \in \mathrm{Max}_{v\in W} P(v | \llbracket f\rrbracket) \mid f \in \sigma_1(w)\}. \tag{3.3}$$

[7] Originally, Blutner (1998) defined preferences in terms of a function C that maps form-meaning pairs to real numbers, where $C(\langle f, w\rangle) = c(f) \times -\log_2 P(w | \llbracket f\rrbracket)$. Form-meaning pairs were then ordered with respect to their C-value. Our formulation here amounts basically to the same, but further integrates the present assumption that costs are nominal and only sender relevant.

Proposition 3.8. For arbitrary interpretation frames it holds that $\rho_1^{\mathrm{OT}}(f) \subseteq \rho_1^{\mathrm{OA}}(f)$. For interpretation frames satisfying strong expressibility it holds that $\rho_1^{\mathrm{OT}}(f) = \rho_1^{\mathrm{OA}}(f)$.

Proof. The first part is an immediate consequences of observation 3.7. So assume strong expressibility and let $\sigma_1^{-1}(f) \neq \emptyset$ and $w \in \rho_1^{\mathrm{OA}}(f)$, so that $f \in \sigma_1(w)$. From observation 3.4 we know that therefore $w \in \rho_0(f)$. So there is no w' for which $P(w' | \llbracket f \rrbracket) > P(w | \llbracket f \rrbracket)$. But that means that $\langle f, w \rangle$ also satisfies the I-principle, and therefore $w \in \rho_1^{\mathrm{OT}}(f)$. Q.E.D.

Proposition 3.9. For arbitrary interpretation frames it holds that $\rho_1^{\mathrm{OT}}(f) \subseteq \rho_1^{\mathrm{BR}}(f)$. For interpretation frames satisfying strong expressibility it holds that $\rho_1^{\mathrm{OT}}(f) = \rho_1^{\mathrm{BR}}(f)$.

Proof. Let $\sigma_1^{-1}(f) \neq \emptyset$ and $w \in \rho_1^{OT}(f)$. Then $w \in \mathrm{Max}_{v \in W} P(v | \llbracket f \rrbracket)$ and $f \in \sigma_1(w)$. Suppose that there was a $w' \in W$ with $P(w' | \sigma_1^{-1}(f)) > P(w | \sigma_1^{-1}(f))$. Then $w' \in \sigma_1^{-1}(f)$, but $w' \notin \llbracket f \rrbracket$. This contradicts observation 3.1. The rest follows from propositions 3.5 and 3.8. Q.E.D.

3.4 Interpretation of Optimal Signals

The results of the last sections are graphically represented in figure 1. What do these results tell us about the respective interpretation rules? In particular, what are the conceptual differences between the approaches? Can we conclude that one is better than the other? A quick glance at equations (3.1), (3.2) and (3.3) reveals that the only difference between frameworks lies in the treatment of probabilities.[8] The optimal assertions approach does not take probabilities into account, iterated best response chooses the most likely interpretations where the received message was optimal and Bi-OT chooses all those most likely interpretations given the semantic meaning of the message where that message was optimal.

The simplest case where predictions differ is where the to be interpreted message f is true in three worlds, $\llbracket f \rrbracket = \{w, v, u\}$, and optimal in two worlds, $\sigma_1^{-1}(f) = \{v, u\}$, with varying degree of probability: $P(w) > P(v) > P(u)$. In this case, the optimal assertions approach selects $\rho_1^{\mathrm{OA}}(f) = \sigma_1^{-1}(f) = \{v, u\}$, iterated best response selects $\rho_1^{\mathrm{BR}}(f) = \{v\}$, while Bi-OT selects $\rho_1^{\mathrm{OT}}(f) = \emptyset$.

This seems to speak for iterated best response, maybe for optimal assertions, but somehow against Bi-OT. On the other hand, we might also credit

[8] Clearly then, for uniform probability distributions strong expressibility collapses into semantic expressibility and all frameworks behave the exact same.

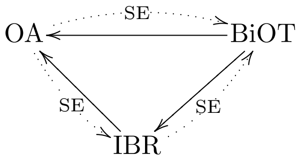

Figure 1. Connection between (i) optimal assertions (OA), (ii) iterated best response (IBR) and (iii) (strong) bidirectional optimality theory (BiOT): a straight arrow indicates inclusion of interpretations of signals while a dotted arrow with label SE indicates inclusion given strong expressibility.

Bi-OT for its strict continuation of the idea that probabilities encode stereotypes in an associative salience ordering: upon hearing f the associations $\rho_0(f)$ spring to mind and those are checked for optimality, so that, if the received message is not optimal in any of the associated worlds in $\rho_0(f)$, then the receiver is stuck — at least for the time being; he might re-associate in a further step.

Can we then make an empirical case for or against any candidate? A first observation is that all three approaches predict the I- and Q-implicature play equally well. In particular, since I- and Q-frames satisfy strong expressibility, the predictions for these cases are exactly the same for all three approaches. The M-frame, on the other hand, does not satisfy strong expressibility, but nevertheless doesn't help judge frameworks, because all of the present candidates mispredict in this case. Take the M-frame as defined above. We then get:

$$\begin{aligned} \rho_1^{\mathrm{OA}}(f) &= \{w, v\} \quad ; \quad \rho_1^{\mathrm{OA}}(g) = \{w, v\} \\ \rho_1^{\mathrm{BR}}(f) &= \{w\} \quad ; \quad \rho_1^{\mathrm{BR}}(g) = \{w, v\} \\ \rho_1^{\mathrm{OT}}(f) &= \{w\} \quad ; \quad \rho_1^{\mathrm{OT}}(g) = \{w, v\} \end{aligned}$$

The problem is that none of the interpretation rules that we considered handles the long form g correctly. Can we fix this problem?

The most obvious idea to try is further iteration. So what would the sender's best response σ_2 be to the receiver's strategy ρ_1? The answer to this question now crucially depends on the uninterpretability assumption 3.2. It is easy to verify that as long as $v \in \rho_1(g)$, the sender's best response will be to send f in w and to send g in v. (Remember that costs are nominal.) To this, in turn, the receiver's best response is the inverse of the sender strategy. The resulting play is indeed the M-implicature play. This is a noteworthy result in the light of the problem of equilibrium selection: iterated best response starting from a 'truth-only' sender strategy *can* account for I-, Q- and M-implicatures for *some* versions of the uninterpretability assumption, but not for others. (To wit, if $\rho_1(g) = \{w\}$ iteration of best responses has reached a fixed-point different from the M-implicature play).

So is the uninterpretability assumption in 3.2 defensible? It does not have to be, since at present it suffices to defend that $\rho_1(g) \neq \{w\}$, which implies that $v \in \rho_1(g)$ as desired. And that $\rho_1(g) \neq \{w\}$ can be argued for based on Cho and Kreps' (1987) intuitive criterion, as has been demonstrated by van Rooij (2007) (see also the short discussion in section 2.3). In simplified terms, the intuitive criterion gives a strong rationale why the receiver should not believe that a sender in w would send g: she has a message f that, given $\rho_1(f)$, is always better in w than signal g *no matter how* the receiver might react to g. (The signal g is *equilibrium-dominated* for w.) This reasoning establishes that $w \notin \rho_1(g)$, which gives us the M-implicature play immediately. If we adopt a weaker version and only require that $\rho_1(g) \neq \{w\}$, we can account for M-implicatures after another round of iteration.

4 Conclusion

Taken together, we may say that, with only little help from the refinement literature, the present version of iterated best response provides a uniform, synchronic account of I-, Q- and M-implicatures. It also subsumes, as a standard game-theoretical model, the optimal assertions approach and strong Bi-OT. This does not discredit either of these latter approaches. For the optimal assertions approach is actually more general than presented here: its predictions were here only assessed for a special case, but the framework is not restricted to a sender who knows the actual world and a receiver who chooses interpretation actions. Similarly, strong optimality is not all there is to Bi-OT: there is also the notion of weak bidirectional optimality which also handles M-implicatures. The connection between weak optimality and iterated best response is not obvious and remains an interesting topic of future research. At present, we may safely conclude that, if game-theoretic standards are a criterion for our selection of models of implicature calculation, then iterated best response fares best in the neo-Gricean terrain.

References

Atlas, J. D. and Levinson, S. (1981). It-clefts, informativeness, and logical-form. In Cole, P., editor, *Radical Pragmatics*, pages 1–61. Academic Press.

Benz, A. and van Rooij, R. (2007). Optimal assertions and what they implicate. *Topoi*, 26:63–78.

Blutner, R. (1998). Lexical pragmatics. *Journal of Semantics*, 15:115–162.

Blutner, R. (2000). Some aspects of optimality in natural language interpretation. *Journal of Semantics*, 17:189–216.

Cho, I.-K. and Kreps, D. M. (1987). Signaling games and stable equilibria. *The Quarterly Journal of Economics*, 102(2):179–221.

Farrell, J. (1993). Meaning and credibility in cheap-talk games. *Games and Economic Behavior*, 5:514–531.

Grice, P. H. (1989). *Studies in the Ways of Words.* Harvard University Press.

Horn, L. R. (1984). Towards a new taxonomy for pragmatic inference: Q-based and I-based implicatures. In Shiffrin, D., editor, *Meaning, Form, and Use in Context*, pages 11–42. Georgetown University Press, Washington.

Jäger, G. (2007). Game dynamics connects semantics and pragmatics. In Pietarinen, A.-V., editor, *Game Theory and Linguistic Meaning*, pages 89–102. Elsevier.

Levinson, S. C. (2000). *Presumptive Meanings. The Theory of Generalized Conversational Implicature.* MIT Press, Cambridge, Massachusetts.

van Rooij, R. (2007). Games and quantity implicature. To appear.