



UvA-DARE (Digital Academic Repository)

Interpretation of optimal signals

Franke, M.

Publication date
2008

Published in
New Perspectives on Games and Interaction

[Link to publication](#)

Citation for published version (APA):

Franke, M. (2008). Interpretation of optimal signals. In K. R. Apt, & R. van Rooij (Eds.), *New Perspectives on Games and Interaction* (pp. 297-310). (Texts in logic and games; Vol. 4). Amsterdam University Press.
http://student.science.uva.nl/~mfranke/Papers/Franke_Interpretation_Optimal_Signals.pdf

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

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 series of iterated best responses (cf. 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-theoretic models of conversational implicature calculation, and this study belongs to this line of research. It provides a formal comparison of selected previous approaches 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's 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).

0087 (4) Sue smiled.

0088 M-implicatures were also discussed by Horn (1984) and have been addressed
 0089 as *Horn's division of pragmatic labor* thereafter. It has become customary
 0090 to assume that both sentences (3) and (4) are semantically equivalent, but,
 0091 when put to use, the longer form (3) gets to be associated with the non-
 0092 stereotypical situation, while the short form (4) gets to be associated with
 0093 the stereotypical situation.
 0094

0095 2 Implicatures via signaling games

0096 2.1 Interpretation frames

0097 A fairly manageable set of contextual parameters plays a role in the neo-
 0098 Gricean classification of implicatures: we distinguish various meanings that
 0099 are more or less stereotypical and we compare different forms with respect
 0100 to their semantic meaning and complexity. We can then capture any such
 0101 configuration of contextual parameters that are relevant for the computation
 0102 of implicatures in an *interpretation frame*.
 0103

0104 **Definition 2.1** (Interpretation Frame). An interpretation frame is a tuple
 0105

$$0106 \mathcal{F} \stackrel{\text{def}}{=} \langle W, P, F, c, \llbracket \cdot \rrbracket \rangle$$

0107 where W is a finite set of worlds or situations, P is a probability distribution
 0108 over W with the usual properties,¹ F is a set of forms or signals which the
 0109 sender may send, $c : F \rightarrow \mathbb{R}$ is a cost function and $\llbracket \cdot \rrbracket : F \rightarrow \mathcal{P}(W)$ is a
 0110 semantic denotation function mapping forms to subsets of W .
 0111

0112 We assume for convenience that $P(w) \neq 0$ for all worlds $w \in W$. We
 0113 would also like to rule out certain rather pathological situations where there
 0114 are worlds which simply cannot be expressed by any conventional signal:
 0115

0116 **Assumption 2.2** (Semantic Expressibility). We only consider interpreta-
 0117 tion frames in which all worlds are semantically expressible: for all worlds
 0118 w there has to be a form f such that $w \in \llbracket f \rrbracket$.
 0119

0120 The kinds of implicatures described in the previous section correspond
 0121 to abstract interpretation frames as follows:
 0122

- 0123 • The *I-frame* is an interpretation frame $\mathcal{F}_I = \langle W, P, F, c, \llbracket \cdot \rrbracket \rangle$ where
 0124 $W = \{w, v\}$, $P(w) > P(v) \neq 0$, $F = \{f, g, h\}$, $c(f) < c(g), c(h)$ and
 0125 $\llbracket f \rrbracket = W$, $\llbracket g \rrbracket = \{v\}$ and $\llbracket h \rrbracket = \{w\}$. The observed *I-implicature play*
 0126 is to interpret f as w and to send f in w only.
 0127

0128 ¹ $P(w) \in [0, 1]$, for all $w \in W$; $P(A) = \sum_{w \in A} P(w)$, for all $A \subseteq W$; $P(W) = 1$.
 0129

- 0130 • The Q -frame is an interpretation frame $\mathcal{F}_Q = \langle W, P, F, c, \llbracket \cdot \rrbracket \rangle$ where
 0131 $W = \{w, v\}$, $P(w) \geq P(v) \neq 0$, $F = \{f, g\}$, $c(f) = c(g)$ and $\llbracket f \rrbracket = W$,
 0132 $\llbracket g \rrbracket = \{v\}$. The observed Q -implicature play is to interpret f as w and
 0133 to send f in w only.
- 0134 • The M -frame is an interpretation frame $\mathcal{F}_M = \langle W, P, F, c, \llbracket \cdot \rrbracket \rangle$ where
 0135 $W = \{w, v\}$, $P(w) > P(v) \neq 0$, $F = \{f, g\}$, $c(f) < c(g)$ and $\llbracket f \rrbracket =$
 0136 $\llbracket g \rrbracket = W$. The observed M -implicature play is to interpret f as w and
 0137 to send f in w only, as well as to interpret g as v and to send g in v
 0138 only.
 0139

0140 2.2 Interpretation games

0141 Interpretation frames capture the relevant aspects of the situation in which
 0142 communication takes place. The communication itself can best be imagined
 0143 as a signaling game: nature selects a world $w \in W$ —call it the actual world
 0144 in a given play—with probability $P(w)$ and reveals it to the sender who
 0145 in turn chooses a form $f \in F$. The receiver does not observe the actual
 0146 world, but observes the signal f . He then chooses an action A . Sender and
 0147 receiver receive a payoff based on w , f and A . In the present context, we are
 0148 interested in *interpretation games*: signaling games in which signals have
 0149 a conventional, compelling meaning that the receiver tries to interpret by
 0150 choosing an interpretation action $\emptyset \neq A \subseteq W$.

0152 **Definition 2.3** (Interpretation Game). An interpretation game is just an
 0153 interpretation frame to which interpretation actions and utilities for sender
 0154 and receiver are added, in other words a tuple

$$0155 \mathcal{G} \stackrel{\text{def}}{=} \langle \mathcal{F}, \text{Act}, u_S, u_R \rangle$$

0157 where $\mathcal{F} = \langle W, P, F, c, \llbracket \cdot \rrbracket \rangle$ is an interpretation frame, $\text{Act} \stackrel{\text{def}}{=} \mathcal{P}(W) \setminus \emptyset$ is a
 0158 set of *interpretation actions* and $u_x : F \times \text{Act} \times W \rightarrow \mathbb{R}$ are utility functions
 0159 of sender and receiver:²
 0160

$$0161 u_R(f, A, w) \stackrel{\text{def}}{=} \begin{cases} \frac{1}{|A|} & \text{if } w \in A \text{ and } w \in \llbracket f \rrbracket \\ 0 & \text{if } w \notin A \text{ and } w \in \llbracket f \rrbracket \\ -1 & \text{otherwise} \end{cases}$$

$$0165 u_S(f, A, w) \stackrel{\text{def}}{=} u_R(f, A, w) - c(f).$$

0167 ² These utilities reflect the mutual desire to communicate which world is actual: the
 0168 more the receiver narrows down a correct guess the better; miscommunication, on the
 0169 other hand, is penalized so that if the chosen interpretation does not include the actual
 0170 situation, the payoff is strictly smaller than when it does; a strong penalty is given
 0171 for communication that deviates from the semantic meaning of messages to enforce
 0172 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.)

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 $\sigma : W \rightarrow \mathcal{P}(F) \setminus \emptyset$ 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 w and f whenever $f \in \sigma(w)$ we have $w \in \llbracket f \rrbracket$. We define also $\sigma^{-1}(f) \stackrel{\text{def}}{=} \{w \in W \mid f \in \sigma(w)\}$. Finally, a *receiver strategy* is a function $\rho : F \rightarrow \text{Act}$ 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 $\rho(f)$:

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

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

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

For a truth-respecting sender strategy this simplifies to:

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

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 Equation 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

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

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 $\text{EU}_R(A, \sigma, f) = 0$ is constant for all A if $\sigma^{-1}(f) = \emptyset$.

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

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 $\text{Max}_{w \in W} P(w|\sigma^{-1}(f))$. 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 $w \in W$: $\sigma(w) \in \text{Max}_{f \in F} \text{EU}_S(f, \rho, w)$
- (ii) for all $f \in F$: $\rho(f) \in \text{Max}_{A \in \text{Act}} \text{EU}_R(A, \sigma, f)$.

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 (2008) 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's 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

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

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* (cf. 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’s 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:

$$\sigma_0(w) = \{f \in F \mid w \in \llbracket f \rrbracket\}.$$

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):

$$\begin{aligned} \rho_0(f) &= \text{Max}_{w \in W} P(w \mid \sigma_0^{-1}(f)) \\ &= \text{Max}_{w \in W} P(w \mid \llbracket f \rrbracket). \end{aligned}$$

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 $\llbracket \cdot \rrbracket$ and $P(\cdot)$: receipt of message f is treated as if it was the observation of the event $\llbracket f \rrbracket$. 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

⁶ An anonymous reviewer asks for the difference between Jäger’s (2007) evolutionary

0302 But if this truly is the most reasonable beginning for pragmatic inter-
 0303 pretation, the sender may anticipate the receiver's association response ρ_0
 0304 and choose a best response to it:

$$\begin{aligned} 0305 \sigma_1(w) &= \text{Max}_{f \in F} \text{EU}_S(f, \rho_0, w) \\ 0306 &= \{f \in F \mid \neg \exists f' \in F : \text{EU}_S(f, \rho_0, w) < \text{EU}_S(f', \rho_0, w)\} \end{aligned}$$

0308 Forms in σ_1 are optimal forms given the receiver's association response. We
 0309 could therefore call them *association optimal*, or, for short, *optimal*: a form
 0310 $f \in F$ is (association) optimal in a world w iff $f \in \sigma_1(w)$.

0311 How should the receiver interpret an optimal signal? We'll next consider
 0312 and compare three possible answers to this question.

0314 3.2 Optimal assertions and iterated best response

0315 Given semantic expressibility as stated in Assumption 2.2, association op-
 0316 timality is equivalent to Benz's and van Rooij's (2007) notion of an optimal
 0317 assertion. Although the latter notion requires truth of a message for its op-
 0318 timality, it is easy to see that semantic expressibility and optimality entail
 0319 truth.

0320 **Observation 3.1.** Given semantic expressibility, σ_1 is truth-respecting.

0321 *Proof.* Let some $f \in F$ be false in $w \in W$. From semantic express-
 0322 ibility there is a message $f' \in F$ which is true in w . But then $-1 =$
 0323 $u_S(f, \rho_0(f), w) < 0 \leq u_S(f', \rho_0(f'), w)$, so that f is not association optimal
 0324 in w . Q.E.D.

0326 If the sender sends an association optimal signal, i.e. if the sender sticks
 0327 to σ_1 , the receiver can again interpret accordingly. Benz and van Rooij
 0328 propose the following interpretation rule based on the assumption that the
 0329 received signal was an *Optimal Assertion*: $\rho_1^{\text{OA}}(f) = \{w \in \llbracket f \rrbracket \mid f \text{ is optimal}$
 0330 $\text{in } w\}$. This simplifies under Observation 3.1 to

$$0332 \rho_1^{\text{OA}}(f) = \sigma_1^{-1}(f). \tag{3.1}$$

0334 Notice, however, that this may not be a well-defined receiver strategy in
 0335 our present set-up, for it may be the case that $\sigma_1^{-1}(f) = \emptyset$, which is not
 0336 a feasible interpretation action. The same problem also occurs for the best
 0337 response to σ_1 . It is clear what the best response to σ_1 is for messages that
 0338 may be optimal somewhere: if $\sigma_1^{-1}(f) \neq \emptyset$, we have

$$0339 \rho_1^{\text{BR}}(f) = \text{Max}_{w \in W} P(w \mid \sigma_1^{-1}(f)). \tag{3.2}$$

0340 model, which also uses best response dynamics, and the present synchronic approach.
 0341 One obvious difference is that the present model assumes that at each turn a best
 0342 response is selected with probability 1. Another difference is the starting point: in
 0343 Jäger's model it is the sender, while in the present model it is receiver who responds
 0344 first to a strategy that is given by the semantic meaning of the signals.

0345 But how should a best response to σ_1 interpret messages that are never
 0346 optimal? Since we defined (tentatively, in Footnote 3) expected utilities
 0347 as constant for all $A \in Act$ whenever $\sigma^{-1}(f) = \emptyset$, any $A \in Act$ is an
 0348 equally good interpretation for a non-optimal f . For our present purpose—
 0349 the comparison of frameworks—it is not important what to choose in this
 0350 case, as long as we choose consistently. We therefore adopt the following
 0351 assumption and reflect on it in Section 3.4 where it plays a crucial role.

0352 **Assumption 3.2** (Uninterpretability Assumption). We assume that the
 0353 receiver resorts to the mere semantic meaning in case a message is uninter-
 0354 pretable: if $\sigma_1^{-1}(f) = \emptyset$, then $\rho_1^{\text{OA}}(f) = \rho_1^{\text{BR}}(f) = \llbracket f \rrbracket$.

0356 With this we can show that $\rho_1^{\text{BR}}(f)$ entails $\rho_1^{\text{OA}}(f)$ for arbitrary f and
 0357 interpretation frames. Moreover, ρ_1^{OA} also entails ρ_1^{BR} , if we assume *strong*
 0358 *expressibility*:

0360 **Definition 3.3** (Strong Expressibility). An interpretation frame satisfies
 0361 strong expressibility if each world is immediately associated with some mes-
 0362 sage: for each world w there is a form f such that $w \in \rho_0(f)$.

0363 **Observation 3.4.** Under strong expressibility, association optimality im-
 0364 plies inclusion in the association response: if f is association optimal in w ,
 0365 then $w \in \rho_0(f)$.

0367 *Proof.* Assume strong expressibility. If $w \notin \rho_0(f)$, there is a form f' for
 0368 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
 0369 not association optimal in w . Q.E.D.

0371 **Proposition 3.5.** For arbitrary interpretation frames it holds that $\rho_1^{\text{BR}}(f)$
 0372 $\subseteq \rho_1^{\text{OA}}(f)$. For interpretation frames satisfying strong expressibility it holds
 0373 that $\rho_1^{\text{BR}}(f) = \rho_1^{\text{OA}}(f)$.

0374 *Proof.* We only have to look at the non-trivial case where $\sigma_1^{-1}(f) \neq \emptyset$. Let
 0375 $w \in \rho_1^{\text{BR}}(f)$. Since all worlds have non-zero probabilities we can conclude
 0376 that $w \in \sigma_1^{-1}(f)$. Hence, $w \in \rho_1^{\text{OA}}(f)$.

0378 Let $w \in \rho_1^{\text{OA}}(f)$ and assume strong expressibility. Then $w \in \llbracket f \rrbracket$ and
 0379 $f \in \sigma_1(w)$. From Observation 3.4 we then know that $w \in \rho_0(f)$. That
 0380 means that there is no w' for which $P(w' | \llbracket f \rrbracket) > P(w | \llbracket f \rrbracket)$. But since, by
 0381 Observation 3.1, we know that $\sigma_1^{-1}(f) \subseteq \llbracket f \rrbracket$, we also know that there is no
 0382 w' for which $P(w' | \sigma_1^{-1}(f)) > P(w | \sigma_1^{-1}(f))$. Hence $w \in \rho_1^{\text{BR}}(f)$. Q.E.D.

0383 3.3 Strong bidirectional optimality theory

0384 A similar connection holds with strong Bi-OT (Blutner, 1998, 2000). At first
 0385 sight, Bi-OT looks rather different from game-theoretic models, because in
 0386 Bi-OT we compare form-meaning pairs $\langle f, w \rangle$ with respect to a preference
 0387

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:

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

This simplifies to:⁷

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

Observation 3.7. Interpretation based on optimal assertions $\rho_1^{\text{OA}}(f)$ 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 $w \in \rho_1^{\text{OA}}(f)$.

Proof. A form-meaning pair $\langle f, w \rangle$ satisfies the Q principle iff there is no f' such that $\text{EU}_S(f', \rho_0, w) > \text{EU}_S(f, \rho_0, w)$ iff f is association optimal in w iff $\sigma_1^{-1}(f) \neq \emptyset$ and $w \in \rho_1^{\text{OA}}(f)$. 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 $\rho_1^{\text{OT}}(f) = \llbracket f \rrbracket$; otherwise: $\rho_1^{\text{OT}}(f) = \{w \in W \mid \langle f, w \rangle \text{ is strongly optimal}\}$, which is equivalent to:

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

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.

0431 **Proposition 3.8.** For arbitrary interpretation frames it holds that $\rho_1^{\text{OT}}(f)$
 0432 $\subseteq \rho_1^{\text{OA}}(f)$. For interpretation frames satisfying strong expressibility it holds
 0433 that $\rho_1^{\text{OT}}(f) = \rho_1^{\text{OA}}(f)$.

0434 *Proof.* The first part is an immediate consequences of Observation 3.7. So
 0435 assume strong expressibility and let $\sigma_1^{-1}(f) \neq \emptyset$ and $w \in \rho_1^{\text{OA}}(f)$, so that
 0436 $f \in \sigma_1(w)$. From Observation 3.4 we know that therefore $w \in \rho_0(f)$. So
 0437 there is no w' for which $P(w' | \llbracket f \rrbracket) > P(w | \llbracket f \rrbracket)$. But that means that $\langle f, w \rangle$
 0438 also satisfies the I-principle, and therefore $w \in \rho_1^{\text{OT}}(f)$. Q.E.D.

0440 **Proposition 3.9.** For arbitrary interpretation frames it holds that $\rho_1^{\text{OT}}(f)$
 0441 $\subseteq \rho_1^{\text{BR}}(f)$. For interpretation frames satisfying strong expressibility it holds
 0442 that $\rho_1^{\text{OT}}(f) = \rho_1^{\text{BR}}(f)$.

0443 *Proof.* Let $\sigma_1^{-1}(f) \neq \emptyset$ and $w \in \rho_1^{\text{OT}}(f)$. Then $w \in \text{Max}_{v \in W} P(v | \llbracket f \rrbracket)$
 0444 and $f \in \sigma_1(w)$. Suppose that there was a $w' \in W$ with $P(w' | \sigma_1^{-1}(f)) >$
 0445 $P(w | \sigma_1^{-1}(f))$. Then $w' \in \sigma_1^{-1}(f)$, but $w' \notin \llbracket f \rrbracket$. This contradicts Observa-
 0446 tion 3.1. The rest follows from Propositions 3.5 and 3.8. Q.E.D.

0448 3.4 Interpretation of optimal signals

0449 The results of the last sections are graphically represented in Figure 1. What
 0450 do these results tell us about the respective interpretation rules? In par-
 0451 ticular, what are the conceptual differences between the approaches? Can
 0452 we conclude that one is better than the other? A quick glance at Equa-
 0453 tions 3.1, 3.2 and 3.3 reveals that the only difference between frameworks
 0454 lies in the treatment of probabilities.⁸ The optimal assertions approach does
 0455 not take probabilities into account, iterated best response chooses the most
 0456 likely interpretations where the received message was optimal and Bi-OT
 0457 chooses all those most likely interpretations given the semantic meaning of
 0458 the message where that message was optimal.

0460 The simplest case where predictions differ is where the to be interpreted
 0461 message f is true in three worlds, $\llbracket f \rrbracket = \{w, v, u\}$, and optimal in two worlds,
 0462 $\sigma_1^{-1}(f) = \{v, u\}$, with varying degree of probability: $P(w) > P(v) > P(u)$.
 0463 In this case, the optimal assertions approach selects $\rho_1^{\text{OA}}(f) = \sigma_1^{-1}(f) =$
 0464 $\{v, u\}$, iterated best response selects $\rho_1^{\text{BR}}(f) = \{v\}$, while Bi-OT selects
 0465 $\rho_1^{\text{OT}}(f) = \emptyset$.

0466 This seems to speak for iterated best response, maybe for optimal as-
 0467 sertions, but somehow against Bi-OT. On the other hand, we might also
 0468 credit Bi-OT for its strict continuation of the idea that probabilities encode
 0469 stereotypes in an associative salience ordering: upon hearing f the associa-
 0470 tions $\rho_0(f)$ spring to mind and those are checked for optimality, so that, if

0471 ⁸ Clearly then, for uniform probability distributions strong expressibility collapses into
 0472 semantic expressibility and all frameworks behave the exact same.

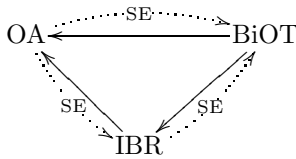


FIGURE 1. Connection between (i) optimal assertions (OA), (ii) iterated best response (IBR) and (iii) (strong) bidirectional optimality theory (BiOT): a straight arrow indicates inclusion of interpretations of signals while a dotted arrow with label SE indicates inclusion given strong expressibility.

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:

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

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,

0517 but not for others. (To wit, if $\rho_1(g) = \{w\}$ iteration of best responses has
 0518 reached a fixed-point different from the M-implicature play).

0519 So is the uninterpretability assumption in 3.2 defensible? It does not
 0520 have to be, since at present it suffices to defend that $\rho_1(g) \neq \{w\}$, which
 0521 implies that $v \in \rho_1(g)$ as desired. And that $\rho_1(g) \neq \{w\}$ can be argued for
 0522 based on Cho's and Kreps' (1987) intuitive criterion, as has been demon-
 0523 strated by van Rooij (2008) (see also the short discussion in Section 2.3).
 0524 In simplified terms, the intuitive criterion gives a strong rationale why the
 0525 receiver should not believe that a sender in w would send g : she has a
 0526 message f that, given $\rho_1(f)$, is always better in w than signal g *no matter*
 0527 *how* the receiver might react to g . (The signal g is *equilibrium-dominated*
 0528 *for* w .) This reasoning establishes that $w \notin \rho_1(g)$, which gives us the M-
 0529 implicature play immediately. If we adopt a weaker version and only require
 0530 that $\rho_1(g) \neq \{w\}$, we can account for M-implicatures after another round
 0531 of iteration.

0532 4 Conclusion

0533
 0534 Taken together, we may say that, with only little help from the refinement
 0535 literature, the present version of iterated best response provides a uniform,
 0536 synchronic account of I-, Q- and M-implicatures. It also subsumes, as a stan-
 0537 dard game-theoretical model, the optimal assertions approach and strong
 0538 Bi-OT. This does not discredit either of these latter approaches. For the
 0539 optimal assertions approach is actually more general than presented here:
 0540 its predictions were here only assessed for a special case, but the framework
 0541 is not restricted to a sender who knows the actual world and a receiver who
 0542 chooses interpretation actions. Similarly, strong optimality is not all there
 0543 is to Bi-OT: there is also the notion of weak bidirectional optimality which
 0544 also handles M-implicatures. The connection between weak optimality and
 0545 iterated best response is not obvious and remains an interesting topic of
 0546 future research. At present, we may safely conclude that, if game-theoretic
 0547 standards are a criterion for our selection of models of implicature calcula-
 0548 tion, then iterated best response fares best in the neo-Gricean terrain.

0549 Acknowledgments

0550
 0551 Thanks to Anton Benz, Reinhard Blutner, Tikitu de Jager, Gerhard Jäger,
 0552 Robert van Rooij and an anonymous referee for very helpful comments and
 0553 discussions.

0554 References

0555
 0556 Atlas, J.D. & Levinson, S. (1981). It-clefts, Informativeness, and Logical-
 0557 Form. In Cole, P., ed., *Radical Pragmatics*, pp. 1–61. Academic Press.
 0558
 0559

- 0560 Benz, A. & van Rooij, R. (2007). Optimal assertions and what they impli-
0561 cate. a uniform game theoretic approach. *Topoi*, 26(1):63–78.
- 0562 Blutner, R. (1998). Lexical Pragmatics. *Journal of Semantics*, 15(2):115–
0563 162.
- 0565 Blutner, R. (2000). Some Aspects of Optimality in Natural Language In-
0566 terpretation. *Journal of Semantics*, 17:189–216.
- 0567
- 0568 Cho, I. & Kreps, D.M. (1987). Signaling Games and Stable Equilibria. *The*
0569 *Quarterly Journal of Economics*, 102(2):179–221.
- 0570
- 0571 Farrell, J. (1993). Meaning and Credibility in Cheap-Talk Games. *Games*
0572 *and Economic Behavior*, 5(4):514–531.
- 0573
- 0574 Grice, H.P. (1989). *Studies in the Way of Words*. Harvard University Press,
0575 Cambridge, Mass.
- 0576
- 0577 Horn, L.R. (1984). Towards a new taxonomy for pragmatic inference: Q-
0578 based and I-based implicatures. In Shiffrin, D., ed., *Meaning, Form, and*
0579 *Use in Context*, pp. 11–42. Georgetown University Press, Washington.
- 0580
- 0581 Jäger, G. (2007). Game dynamics connects semantics and pragmatics. In
0582 Pietarinen, A.-V., ed., *Game Theory and Linguistic Meaning*, pp. 89–102.
Elsevier.
- 0583
- 0584 Levinson, S.C. (2000). *Presumptive Meanings. The Theory of Generalized*
0585 *Conversational Implicature*. MIT Press, Cambridge, Massachusetts.
- 0586
- 0587 van Rooij, R. (2008). Games and Quantity Implicature. *Journal of Eco-*
0588 *nomic Methodology*. To appear.
- 0589
- 0590
- 0591
- 0592
- 0593
- 0594
- 0595
- 0596
- 0597
- 0598
- 0599
- 0600
- 0601
- 0602