Interpretation games with variable costs

Gerhard Jäger
Seminar für Sprachwissenschaft, University of Tübingen, Germany
gerhard.jaeger@uni-tuebingen.de

1 Interpretation games

Game theory has proved a versatile tool to model the Grice-style reasoning processes enabling interlocutors to figure out *what is meant* from *what is said* and contextual information. In this position statement, I will give a brief overview of the *Iterated Best Response* model of game-theoretic pragmatics, a particular incarnateion of this general program that has emerged within the past five years (Franke 2009, 2011; Franke and Jäger 2012; Jäger 2008, 2011, 2012; Jäger and Ebert 2009). In this brief communication, I will rely mainly on the version of the model as devoleped by Michael Franke (2009; 2011) and propose a modification regarding the treatment of message costs.

The contextual information is modeled as an *interpretation game*, a variant of signaling games in the sense of Lewis (1969). It consists of a set of types T, a finite set of of messages M, a set of actions A, a prior probability distribution $\mathbf{p}$ over T (with $\mathbf{p} \in \Delta^+(T)$, i.e. $\mathbf{p}$ assigns a strictly positive probability to each type), a cost function $\mathbf{c} \in M \mapsto [0, \infty)$, two utility functions $u_s, u_r \in T \times A \mapsto \mathbb{R}$, and an interpretation function $\| \cdot \| \in M \mapsto POW(T)$.

A history of a game is a sequence $\langle t, m, a \rangle \in T \times M \times A$. The payoff of sender and receiver are defined over histories as:

$$\begin{aligned} u_s(t, m, a) &= u_s(t, a) - \mathbf{c}(m) \\ u_r(t, m, a) &= u_r(t, a) \end{aligned}$$

A (behavioral) sender strategy is a stochastic function σ from types to messages, i.e. a function from types to probability distributions over messages. Likewise, a receiver strategy is a function ρ from messages to probability distributions over actions.

In this paper we will only consider games where $T = A$ and where both u_s and u_r are Kronecker delta functions, i.e.

$$u_{s/r}(t, a) = \begin{cases} 1 & \text{if } t = a \\ 0 & \text{else} \end{cases}$$

Let us consider a standard example, the emergence of a scalar implicature triggered by the determiner *some*.

(1) a. Joe ate some of the cookies. (m_{some})
b. Joe ate all of the cookies. (m_{all})

	m_{some}	m_{all}	$\mathbf{p}$
$t_{\exists\neg\forall}$	1	0	$1/2$
$t_{\forall}$	1	1	$1/2$
$\mathbf{c}$	0	0	

Table 1: Some-all game

If it is understood by both interlocutors that the listener is interested in how many cookies Joe ate, answer (1a) triggers the implicature that he did not eat all of the cookies.

The corresponding interpretation game can concisely be represented as in Table 1: The first two rows represent the two types $t_{\exists\neg\forall}$ (where Joe ate some but not all of the cookies) and $t_{\forall}$ (where he ate all of them). The first two columns represent the messages ((1a) and (b) respectively). The Boolean entries in the upper left 2×2 sub-table gives the interpretation function. The last row gives the cost function $\mathbf{c}$ and the last column the prior probabilities $\mathbf{p}$.

2 The IBR sequence

Iterated Best Response (IBR) is a protocol to select a strategy profile for such a game via iterated back-and-forth reasoning. The default strategy σ_0 of an honest sender maps each type t to a uniform distribution over all messages m with $t \in \|m\|$. Likewise, the default strategy ρ_0 of a trusting receiver maps each message m to a uniform distribution over all actions a with $a \in \|m\|$.

A rational sender who believes that the listener plays ρ_0 will use a strategy σ_1 that maximizes her expected utility given that belief. The same holds *mutatis mutandis* for a sophisticated receiver who beliefs that the sender plays σ_0. Such players are called *level-1*. A rational player believing that their partner is a level-1 player will play a best response to σ_1/ρ_1, making them level-2 players, etc.

The following definitions formalize this basic idea:

Definition 1 (Best response)

$$BR_s(t,\rho) = \arg\max_{m'} \rho(t|m') - \mathbf{c}(m')$$

$$BR_r(m,\sigma) = \begin{cases} \arg\max_a \sigma(m|a)\mathbf{p}(a) & \textit{if } \max_a \sigma(m|a) > 0 \\ \|m\| & \textit{else} \end{cases}$$

Definition 2 (IBR sequence)

$$\sigma_0(m|t) = \begin{cases} 1/|\{m'|t\in\|m'\|\}| & \textit{if } t \in \|m\| \\ 0 & \textit{else} \end{cases}$$

$$\rho_0(a|m) = \begin{cases} 1/|\{a'|a'\in\|m\|\}| & \textit{if } a \in \|m\| \\ 0 & \textit{else} \end{cases}$$

$$\sigma_{k+1}(m|t) = \begin{cases} 1/|BR_s(t,\rho_k)| & \textit{if } m \in BR_s(t,\rho_k) \\ 0 & \textit{else} \end{cases}$$

$$\rho_{k+1}(a|m) = \begin{cases} 1/|BR_r(m,\sigma_k)| & \textit{if } a \in BR_r(m,\sigma_k) \\ 0 & \textit{else} \end{cases}$$

The IBR sequence for the some-all game converges quickly:

$$\sigma_0 = \begin{array}{c} \\ t_{\exists\neg\forall} \\ t_\forall \end{array} \begin{array}{c} \begin{array}{cc} m_{\text{some}} & m_{\text{all}} \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \end{pmatrix} \end{array} \qquad \rho_0 = \begin{array}{c} \\ m_{\text{some}} \\ m_{\text{all}} \end{array} \begin{array}{c} \begin{array}{cc} t_{\exists\neg\forall} & t_\forall \end{array} \\ \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \end{pmatrix} \end{array}$$

$$\sigma_1 = \begin{array}{c} \\ t_{\exists\neg\forall} \\ t_\forall \end{array} \begin{array}{c} \begin{array}{cc} m_{\text{some}} & m_{\text{all}} \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array} \qquad \rho_1 = \begin{array}{c} \\ m_{\text{some}} \\ m_{\text{all}} \end{array} \begin{array}{c} \begin{array}{cc} t_{\exists\neg\forall} & t_\forall \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}$$

$\sigma_2 = \sigma_1$ $\qquad$ $\rho_2 = \rho_1$

In the fixed point (σ_1, ρ_1), m_{some} induces a scalar implicature, i.e. it is associated exclusively with the type $t_{\exists\neg\forall}$ by both players.

3 Problematic examples

While this model captures a sizeable number of pragmatic inferences correctly, there are a few examples where the IBR sequence does not converge to the intuitively correct equilibrium. Two such games are shown in Table 2.

	m_{some}	m_{all}	m_{sbna}	**p**
$t_{\exists\neg\forall}$	1	0	1	1/2
$t_\forall$	1	1	0	1/2
c	0	0	0.1	

	m_{open}	$m_{\text{open-h}}$	$m_{\text{open-a}}$	**p**
t_h	1	1	0	2/3
t_a	1	0	1	1/3
c	0	0.1	0.1	

Table 2: Some-but-not-all game (left) and I-implicature game (right)

The *some-but-not-all* game (see Franke 2009) is an extension of the some-all game where a third message is admitted:

(2) Joe ate some but not all of the cookies. (m_{sbna})

This message is only true in $t_{\exists\neg\forall}$, and it is more complex than the other two messages. This is captured by assigning it a small cost of 0.1. This leads to the IBR sequence:

$$\sigma_0 = \begin{array}{c} \\ t_{\exists\neg\forall} \\ t_\forall \end{array} \begin{array}{c} \begin{array}{ccc} m_{\text{some}} & m_{\text{all}} & m_{\text{sbna}} \end{array} \\ \begin{pmatrix} 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix} \end{array} \qquad \rho_0 = \begin{array}{c} \\ m_{\text{some}} \\ m_{\text{all}} \\ m_{\text{sbna}} \end{array} \begin{array}{c} \begin{array}{cc} t_{\exists\neg\forall} & t_\forall \end{array} \\ \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array}$$

$$\sigma_1 = \begin{array}{c} \\ t_{\exists\neg\forall} \\ t_\forall \end{array} \begin{array}{c} \begin{array}{ccc} m_{\text{some}} & m_{\text{all}} & m_{\text{sbna}} \end{array} \\ \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \end{array} \qquad \rho_1 = \begin{array}{c} \\ m_{\text{some}} \\ m_{\text{all}} \\ m_{\text{sbna}} \end{array} \begin{array}{c} \begin{array}{cc} t_{\exists\neg\forall} & t_\forall \end{array} \\ \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array}$$

$\sigma_2 = \sigma_1$ $\qquad$ $\rho_2 = \rho_1$

In the fixed point, message m_{some} is never used, and it is interpreted according to its literal meaning. This outcome is odd because we would expect the scalar implicature not to be affected by the additional message.

The second game shown in Table 2 is inspired by the following example:

(3) a. John opened the door. (m_{open})

b. John opened the door with the handle. ($m_{\text{open-h}}$)

c. John opened the door with an axe. ($m_{\text{open-a}}$)

While the literal meaning of (3a) is true both in a scenario where John opened the door with the handle (type t_h) and in a scenario where he uses an axe (t_a), we would normally interpret it in the sense of (3b). This is based on the facts that doors are typically opened rather with handles than with axes, and that the unspecific message (3a) is shorter than the more specific (3b). This type of inference is called *I-implicature* by Levinson (2000).

The basic features of this example are captured by the assumptions that t_h has a higher prior probability than t_a and that m_{open} is less costly than $m_{\text{open-h}}$ or $m_{\text{open-a}}$.

The IBR sequence is

$$\sigma_0 = \begin{array}{c} \\ t_h \\ t_a \end{array} \begin{array}{c} \begin{array}{ccc} m_{\text{open}} & m_{\text{open-h}} & m_{\text{open-a}} \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \end{pmatrix} \end{array} \qquad \rho_0 = \begin{array}{c} \\ m_{\text{open}} \\ m_{\text{open-h}} \\ m_{\text{open-a}} \end{array} \begin{array}{c} \begin{array}{cc} t_h & t_a \end{array} \\ \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}$$

$$\sigma_1 = \begin{array}{c} \\ t_h \\ t_a \end{array} \begin{array}{c} \begin{array}{ccc} m_{\text{open}} & m_{\text{open-h}} & m_{\text{open-a}} \end{array} \\ \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array} \qquad \rho_1 = \begin{array}{c} \\ m_{\text{open}} \\ m_{\text{open-h}} \\ m_{\text{open-a}} \end{array} \begin{array}{c} \begin{array}{cc} t_h & t_a \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}$$

$$\sigma_2 = \begin{array}{c} \\ t_h \\ t_a \end{array} \begin{array}{c} \begin{array}{ccc} m_{\text{open}} & m_{\text{open-h}} & m_{\text{open-a}} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array} \qquad \rho_2 = \begin{array}{c} \\ m_{\text{open}} \\ m_{\text{open-h}} \\ m_{\text{open-a}} \end{array} \begin{array}{c} \begin{array}{cc} t_h & t_a \end{array} \\ \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}$$

$\sigma_3 = \sigma_1$ $\qquad$ $\rho_3 = \rho_1$

This sequence does not converge to a fixed point but keeps alternating between (σ_1, ρ_1) and (σ_2, ρ_2). The intuitively correct strategy profile would be (σ_2, ρ_1) though.

4 Variable costs

The reason for these wrong prediction is similar in both cases. When designing her best response to ρ_0 in the some-but-not-all game in type $t_{\exists\neg\forall}$, the sender has the choice between (a) a cheap message with induces a sub-optimal posterior probability for the correct type and (b) a costly message that does uniquely identify the correct type. As long as costs are nominal, the rational choice is the more specific costly message. If the sender would choose the cheap message with some small probability, the IBR sequence would converge to the correct fixed point. The same applies to the best response to ρ_0 in the I-implicature game.

So far we assumed that message costs are common knowledge. The mentioned problems can be remedied if the receiver does not know the sender's costs for sure in advance. To formalize this intution, we generalize the definition of intepretation games given above in the following way: instead of a vector $\mathbf{c}$ of costs, the specification of a game contains a probability density function $\mathcal{C}$ over $\mathbf{c} \in [0,\infty)^{|M|}$, i.e. over cost vectors $\mathbf{c}$.

The cost density $\mathcal{C}$ is common knowledge between the players while the sender knows the exact cost vector. So technically speaking, t and $\mathbf{c}$ represent the sender's type, and $\mathbf{p}, \mathcal{C}$ jointly define the receiver's prior probability distribution over sender types. As t affects the receiver's payoff while $\mathbf{c}$ does not, it is conceptually justified to keep these two aspects of the game structure separate though.

Regarding the some-but-not-all game, let us assume that $\mathcal{C}$ has the following properties:

$$\int_0^\infty \mathcal{C}(0,0,x)dx = 1$$
$$\int_{1/2}^\infty \mathcal{C}(0,0,x)dx = \alpha > 0$$

In words, we assume that m_{some} and m_{all} are costless with probability 1, and that there is a positive probability α that $\mathbf{c}(m_{\text{sbna}}) \geq 1/2$.

With these assumptions, the IBR sequence now comes out as

$$\sigma_0 = \begin{array}{c} \\ t_{\exists\neg\forall} \\ t_\forall \end{array} \begin{array}{c} \begin{array}{ccc} m_{\text{some}} & m_{\text{all}} & m_{\text{sbna}} \end{array} \\ \begin{pmatrix} 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{pmatrix} \end{array} \qquad \rho_0 = \begin{array}{c} \\ m_{\text{some}} \\ m_{\text{all}} \\ m_{\text{sbna}} \end{array} \begin{array}{c} \begin{array}{cc} t_{\exists\neg\forall} & t_\forall \end{array} \\ \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array}$$

$$\sigma_1 = \begin{array}{c} \\ t_{\exists\neg\forall} \\ t_\forall \end{array} \begin{array}{c} \begin{array}{ccc} m_{\text{some}} & m_{\text{all}} & m_{\text{sbna}} \end{array} \\ \begin{pmatrix} \alpha & 0 & 1-\alpha \\ 0 & 1 & 0 \end{pmatrix} \end{array} \qquad \rho_1 = \begin{array}{c} \\ m_{\text{some}} \\ m_{\text{all}} \\ m_{\text{sbna}} \end{array} \begin{array}{c} \begin{array}{cc} t_{\exists\neg\forall} & t_\forall \end{array} \\ \begin{pmatrix} 1/2 & 1/2 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array}$$

$$\sigma_2 = \begin{array}{c} \\ t_{\exists\neg\forall} \\ t_\forall \end{array} \begin{array}{c} \begin{array}{ccc} m_{\text{some}} & m_{\text{all}} & m_{\text{sbna}} \end{array} \\ \begin{pmatrix} \alpha & 0 & 1-\alpha \\ 0 & 1 & 0 \end{pmatrix} \end{array} \qquad \rho_2 = \begin{array}{c} \\ m_{\text{some}} \\ m_{\text{all}} \\ m_{\text{sbna}} \end{array} \begin{array}{c} \begin{array}{cc} t_{\exists\neg\forall} & t_\forall \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array}$$

$$\sigma_3 = \begin{array}{c} \\ t_{\exists\neg\forall} \\ t_\forall \end{array} \begin{array}{c} \begin{array}{ccc} m_{\text{some}} & m_{\text{all}} & m_{\text{sbna}} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \end{array} \qquad \rho_3 = \begin{array}{c} \\ m_{\text{some}} \\ m_{\text{all}} \\ m_{\text{sbna}} \end{array} \begin{array}{c} \begin{array}{cc} t_{\exists\neg\forall} & t_\forall \end{array} \\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix} \end{array}$$

$\sigma_4 = \sigma_3$ $\qquad$ $\rho_4 = \rho_3$

For the I-implicature game to come out correctly, we demand that

$$\int_0^\infty \int_0^\infty \mathcal{C}(0,x,y)dxdy = 1$$
$$\int_0^\infty \int_{1/2}^\infty \mathcal{C}(0,x,y)dxdy = \alpha > 0$$
$$\int_{1/2}^\infty \int_0^\infty \mathcal{C}(0,x,y)dxdy = \beta$$
$$\alpha \mathbf{p}(t_a) > \beta \mathbf{p}(t_h)$$

This means that $\mathbf{c}(m_{\text{open}}) = 0$ with certainty, the marginal probability that $\mathbf{c}(m_{\text{open-h}}) > 1/2$ is a positive value α, and α is at least half as high as (or, more generally, by at least a factor $\mathbf{p}(t_h)/\mathbf{p}(t_a)$ higher than) the marginal probability that $\mathbf{c}(m_{\text{open-a}}) > 1/2$.

Under these assumptions, we arrive at the IBR sequence

$$\sigma_0 = \begin{array}{c} \\ t_h \\ t_a \end{array} \begin{array}{c} \begin{array}{ccc} m_{\text{open}} & m_{\text{open-h}} & m_{\text{open-a}} \end{array} \\ \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \end{pmatrix} \end{array} \qquad \rho_0 = \begin{array}{c} \\ m_{\text{open}} \\ m_{\text{open-h}} \\ m_{\text{open-a}} \end{array} \begin{array}{c} \begin{array}{cc} t_h & t_a \end{array} \\ \begin{pmatrix} 1/2 & 1/2 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}$$

$$\sigma_1 = \begin{array}{c} \\ t_h \\ t_a \end{array} \begin{array}{c} \begin{array}{ccc} m_{\text{open}} & m_{\text{open-h}} & m_{\text{open-a}} \end{array} \\ \begin{pmatrix} \alpha & 1-\alpha & 0 \\ \beta & 0 & 1-\beta \end{pmatrix} \end{array} \qquad \rho_1 = \begin{array}{c} \\ m_{\text{open}} \\ m_{\text{open-h}} \\ m_{\text{open-a}} \end{array} \begin{array}{c} \begin{array}{cc} t_h & t_a \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}$$

$$\sigma_2 = \begin{array}{c} \\ t_h \\ t_a \end{array} \begin{array}{c} \begin{array}{ccc} m_{\text{open}} & m_{\text{open-h}} & m_{\text{open-a}} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array} \qquad \rho_2 = \begin{array}{c} \\ m_{\text{open}} \\ m_{\text{open-h}} \\ m_{\text{open-a}} \end{array} \begin{array}{c} \begin{array}{cc} t_h & t_a \end{array} \\ \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix} \end{array}$$

$\sigma_3 = \sigma_2$ $\qquad$ $\rho_3 = \rho_2$

5 Summary of main claims

1. As spelled out in more detail in the work cited above, the IBR model of game theoretic pragmatics provides a comprehensive formalization of the neo-Gricean program. It correctly captures a wide range of pragmatic inferences. These include

   - scalar implicatures,
   - free choice readings,
   - ignorance implicatures,
   - Horn's division of pragmatic labor,
   - the pragmatics of measure phrases, and
   - putative embedded scalar implicatures.

2. In its standard formulation, the IBR model systematically leads to wrong predictions in scenarios where a cheap non-specific message competes with a more specific costly message. The common assumption of nominal costs implies that a rational sender will always opt for the costly messages in such a case. Assuming non-nominal costs would solve the problem but appears to be an *ad hoc* decision. In this brief paper I argued that the problem can be overcome if we give up the assumption that message costs are common knowledge. As long as the receiver cannot exclude with certainty that costs are non-nominal, IBR converges to the intuitively correct fixed point in the problematic examples.

References

Franke, M. (2009). *Signal to Act: Game Theory in Pragmatics*. Ph.D. thesis, University of Amsterdam.

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, **4**(1):1–82.

Franke, M. and G. Jäger (2012). Bidirectional optimization from reasoning and learning in games. *Journal of Logic, Language and Information*, **21**(1):117–139.

Jäger, G. (2008). Applications of game theory in linguistics. *Language and Linguistics Compass*, **2/3**:408–421.

Jäger, G. (2011). Game-theoretical pragmatics. In J. van Benthem and A. ter Meulen, eds., *Handbook of Logic and Language*, pp. 467–491. Elsevier, 2nd edition.

Jäger, G. (2012). Game theory in semantics and pragmatics. In C. Maienborn, P. Portner, and K. von Heusinger, eds., *Semantics. An International Handbook of Natural Language Meaning*, volume 3, pp. 2487–2516. de Gruyter, Berlin.

Jäger, G. and C. Ebert (2009). Pragmatic rationalizability. In A. Riester and T. Solstad, eds., *Proceedings of Sinn und Bedeutung 13*, number 5 in SinSpeC. Working Papers of the SFB 732, pp. 1–15. University of Stuttgart.

Levinson, S. C. (2000). *Presumptive Meanings*. MIT Press.

Lewis, D. (1969). *Convention*. Harvard University Press, Cambridge, Mass.