Let's pretend to agree

A game theoretic reconstruction of M-implicatures

Gerhard Jäger

Gerhard.Jaeger@uni-bielefeld.de

DIALOR'05

Overview

- signaling games
- saying and meaning in cheap talk signaling games
- if talk is not cheap ...
- Q, I and M
- generalized conventions
- conclusion

Signaling games

- sequential game:
  1. **nature** chooses a world w
     - out of a pool of possible worlds W
     - according to a certain probability distribution P
  2. nature shows w to sender **S**
  3. S chooses a signal/form f out of a set of possible signals F
  4. S transmits f to the receiver **R**
  5. R guesses a meaning $m \in M$

Signaling games

- utility of either player depends both on w and on m
- *cheap talk*: utility does not depend on f
- interests of S and R need not coincide

Signaling games: an example

Example (from Stalnaker 2006):

	m_1	m_2	m_3	m_4
w_1	5 / 5	10 / 10	0 / 0	0 / 0
w_2	5 / 5	0 / 0	6 / 0	8 / 1
w_3	5 / 5	0 / 0	6 / 6	0 / 0

(Each cell: top right / bottom left)

rows: *worlds*

columns: *meanings*

bottom left: *S's utility*

top right: *R's utility*

Stalnaker's example (cont.)

Suppose

- $p(w_1) = P(w_2) = P(w_3) = \frac{1}{3}$
- there are four signals
- signals have the "conventional meanings" $\{w_1\}, \{w_2\}, \{w_3\}$, and $\{w_1, w_2, w_3\}$

Stalnaker's example (cont.)

naïve R:

$$R: \begin{bmatrix} \{w_1\} & \rightarrow & m_2 \\ \{w_2\} & \rightarrow & m_4 \\ \{w_3\} & \rightarrow & m_3 \\ W & \rightarrow & m_1 \end{bmatrix}$$

Stalnaker's example (cont.)

best response of S:

$$R : \begin{bmatrix} \{w_1\} & \rightarrow & m_2 \\ \{w_2\} & \rightarrow & m_4 \\ \{w_3\} & \rightarrow & m_3 \\ W & \rightarrow & m_1 \end{bmatrix}$$

$$S : \begin{bmatrix} w_1 & \rightarrow & \{w_1\} \\ w_2 & \rightarrow & W \\ w_3 & \rightarrow & \{w_3\} \end{bmatrix}$$

Stalnaker’s example (cont.)

best response of R:

$$S : \begin{bmatrix} w_1 & \longrightarrow & \{w_1\} \\ w_2 & \longrightarrow & W \\ w_3 & \longrightarrow & \{w_3\} \end{bmatrix} \quad \longrightarrow \quad R : \begin{bmatrix} \{w_1\} & \longrightarrow & m_2 \\ \{w_2\} & \longrightarrow & ? \\ \{w_3\} & \longrightarrow & m_3 \\ W & \longrightarrow & m_4 \end{bmatrix}$$

Stalnaker's example (cont.)

best response of S:

$$S : \begin{bmatrix} w_1 & \longrightarrow & \{w_1\} \\ w_2 & \longrightarrow & W \\ w_3 & \longrightarrow & \{w_3\} \end{bmatrix} \longleftarrow R : \begin{bmatrix} \{w_1\} & \longrightarrow & m_2 \\ \{w_2\} & \longrightarrow & ? \\ \{w_3\} & \longrightarrow & m_3 \\ W & \longrightarrow & m_4 \end{bmatrix}$$

Some observations

- fixed point of *iterated best response* is Nash equilibrium
- R effectively interprets the signal with the literal meaning W—the tautology—as $\{w_2\}$
- strengthening from W to $\{w_2\}$ can be considered an **implicature**
- schematically:
  - starting point: **semantics**
  - fixed point of iterated best response: **pragmatics**

Cooperative games

- I will restrict attention to games where interests of S and R coincide:

$$u_S = u_R$$

- common goal is the efficient transmission of information:

$$M = POW(W)$$

- "nature's" probability distribution P is assumed to be common knowledge
- utility can thus be defined as

$$u(w, m) = P(w|m)$$

Costly signaling

- talk is not cheap
  - complexity of signals are costs (= negative utility)
  - signals differ in complexity
- $c(f)$: costs (positive real number)
- utility in world w of signal f which is interpreted as meaning m:

$$P(w|m) - c(f)$$

Utility of strategies

- overall utility is determined by **strategies**
  - sender strategy: function $S : W \mapsto F$
  - receiver strategy: function $R : F \mapsto POW(W)$
  - average utility (depends on nature's probability function):

$$u_P(S, R) = \sum_{w \in W} P(w) \cdot (P(w|R(S(w))) - c(S(w)))$$

Implicatures

- Levinson (2000): three types of implicatures
  - Q-implicatures
  - I-implicatures
  - M-implicatures
- all three types of implicatures can be shown to follow from iterated best response under natural assumptions on costs and probabilities

Implicatures

The Q-Heuristics

“What isn’t said, isn’t.”

- related to Grice’s Maxim of Quantity
- accounts for scalar and clausal implicatures

(1) a. Some boys came in. $\rightsquigarrow$ Not all boys came in.
b. Three boys came in. $\rightsquigarrow$ Exactly three boys came in.

(2) a. If John comes, I will leave. $\rightsquigarrow$ It is open whether John comes.
b. John tried to reach the summit. $\rightsquigarrow$ John did not reach the summit.

Q-implicatures

(B = boy, C = come in)

- worlds
  - $w_1 : \exists x.Bx \wedge \forall y.By \longrightarrow Cy$
  - $w_2 : \exists x.Bx \wedge Cx \wedge \exists y.By \wedge \neg Cy$
  - $w_3 : \exists x.Bx \wedge \neg \exists y.By \wedge Cy$
- probabilities

$$P_i(t_1) = P_i(t_2) = P_i(t_3) = 0.3333$$

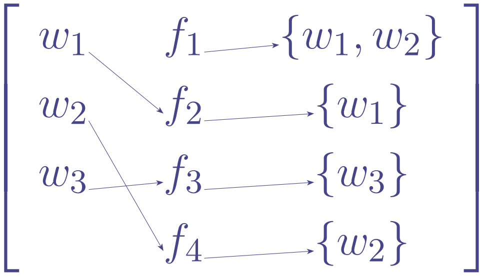
Q-implicatures

- signals:
  - f_1: “Some boys came in.”
  - f_2: “All boys came in.”
  - f_3: “No boys came in.”
  - f_4: “Some, but not all boys came in.”
- costs:

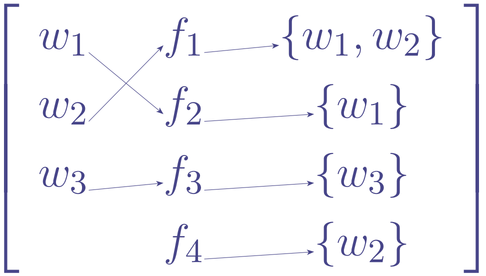
$$c(m_1) = c(m_2) = c(m_3) < c(m_4) - 0.5$$

Q-implicatures

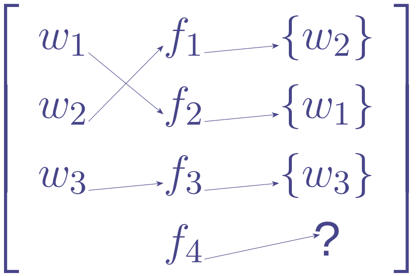
1. semantic convention:



2. *Best response* of S:



3. *Best response* von R:



Q-implicatures

- one round of best response on each side leads to a fixed point
- justifies the (Q-)implicature

  "Some boys came in." *implicates* $\exists x.Bx \wedge \neg Cx$

Q-implicatures

- essentially by Gricean reasoning:
  - there are two competing expressions of similar complexity
  - the literal meaning of the first expression entails the literal meaning of the second expression
  - the speaker wants the hearer to be as well-informed as possible
  - hence the weaker expression can only be used if the stronger one is false
  - hence the stronger expression implicates that the weaker expression is false

I-implicatures

The I-Heuristics

"What is expressed simply is stereotypically exemplified."

- related to Maxim of Manner
- accounts for
  - pragmatic strengthening

(3) a. John's book is good. $\rightsquigarrow$ The book that John is reading or that he has written is good.
b. a secretary $\rightsquigarrow$ a female secretary
c. road $\rightsquigarrow$ hard-surfaced road

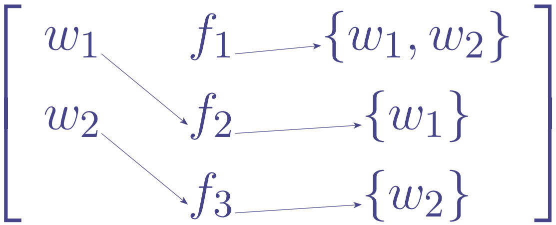
  - ...

I-implicatures

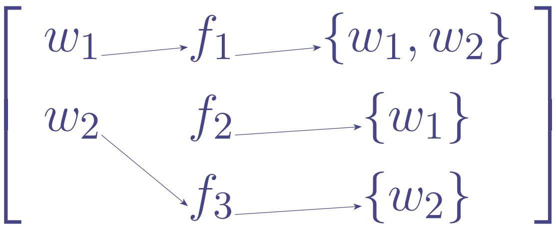
- worlds:
  - w_1: hard-surfaced road
  - w_2: soft-surfaced road
- probabilities
  - $P(w_1) \gg P(w_2)$
  - lets say: $P(w_1) = 9 \cdot P(w_2)$
- signals:
  - f_1: “road”
  - f_2: “hard-surfaced road”
  - f_3: “soft-surfaced road”
- costs:
  - $c(f_1) = 0.10$
  - $c(f_2) = 0.25$
  - $c(f_3) = 0.25$

I-implicatures

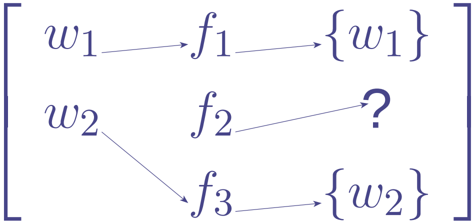
1. semantic convention:



2. *Best response* of S:



3. *Best response* of R:



I-implicatures

- conflicting interests for the speaker:
  - incentive to avoid costs (Manner): use f_1 in w_1
  - incentive to maximize information (Quantity): use f_2 in w_1
- depending on concrete probabilities and costs, either motivation may be stronger
- however: if Manner wins over Quantity, it will always be the more probable (“stereotypical”) denotation that is implicated

M-implicatures

The M-heuristics

> “What is said in an abnormal way isn’t normal.”

(4) a. Bill stopped the car. $\rightsquigarrow$ He used the foot brake.
b. Bill caused the car to stop. $\rightsquigarrow$ He did it in an unconventional way. (like using the hand brake or by making a sharp u-turn)

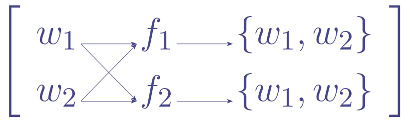
(5) a. Sue smiled. $\rightsquigarrow$ Sue smiled in a regular way.
b. Sue lifted the corners of her lips. $\rightsquigarrow$ Sue produced an artificial smile.

M-implicatures

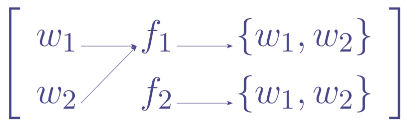
- worlds:
  - w_1: to smile genuinely
  - w_2: to lift the corners of the lips without real smiling
- probabiliites
  - $P_i(w_1) = 9 \cdot P_i(w_2)$
- signals:
  - f_1: “to smile”
  - f_2: “to lift the corner of the lips”
- costs
  - $c(f_1) < c(f_2) - 0.1$

M-implicatures

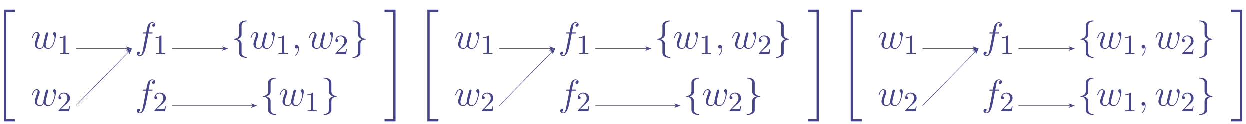
1. semantic convention:



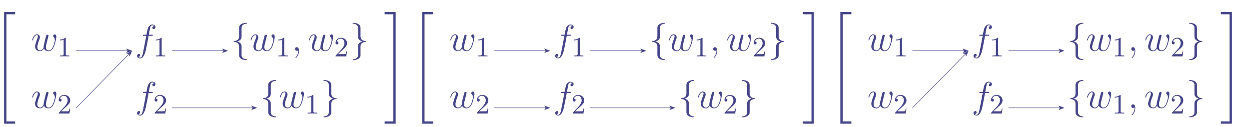
2. *best response* of S:



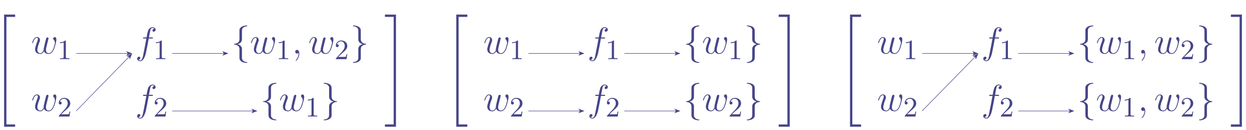
3. *best responses* of R:



4. *best responses* of S:



5. *best response* of R:



M-implicatures

- *best response* is non-deterministic; there may be several best responses
- in above example, three differnt fixed points can be reached via *iterated best response*
- two of them are (non-strict) **pooling equilibria**: no correlations between world and signal
- one (strict) **separating equilibrium**: 1-1 correspondence between world and signal
- this separating equilibrium realizes the M-implicature

  “to lift the corner of the lips” implicates *artificial smile*

M-implicatures

- *Didn't you tune up the parameters to make this work?*

M-implicatures

- yes and no; here is the general pattern:
  - if
    $$|c(f_1) - c(f_2)| > \max(P(w_1), P(w_2))$$
    only reachable fixed point is a pooling equilibrium $\rightsquigarrow$ no implicatures arise
  - if
    $$\min(P(w_1), P(w_2)) < |c(f_1) - c(f_2)| \leq \max(P(w_1), P(w_2))$$
    only reachable strict fixed point is separating equilibrium: cheap signal is assigned to probable meaning and expensive signal to improbable meaning $\rightsquigarrow$ M-implicature

M-implicatures

- if

$$\min(P(w_1), P(w_2)) \geq |c(f_1) - c(f_2)|$$

  both separating equilibria are reachable via iterated best response $\leadsto$ no implicature can be computed

If the parameters are so that they lead to a unique strict equilibrium under iterated best response, this equilibrium realizes the M-implicature.

Generalized conventions

- *Convention* according to Lewis:
  - coordination problem (cooperative game with at least two strict Nash equilibria)
  - Nash equilibrium c
  - common knowledge between players, that everybody plays c

Generalized conventions

- can be generalized
  - hearer believes that it is common knowledge that Santa Claus exists, or
  - speaker believes that hearer believes that it is common knowledge that Santa Claus exists, or
  - hearer believes speaker believes that hearer believes that it is common knowledge that Santa Claus exists, or
  - ...

Generalized conventions

Definition 1 *φ is **a convention for** A **between** A **and** B iff*

1. *ψ is the weakest proposition such that:*

$$\psi \equiv B_A(\varphi \wedge \psi) \wedge B_B(\varphi \wedge \psi)$$

2. *for some n: $B_A B_{i_1} B_{i_2} \cdots B_{i_n} \psi$, where $i_k = A$ for even k and $i_k = B$ for odd k.*

Intuition: φ is a convention for A if it makes sense for A to pretend that φ is common knowledge between A and B.

Conventions and iterated best response

Theorem 1 *Let S and R be the players in a two-person game, and $c = \langle S, R \rangle$ be a convention for S and R between S and R. Suppose that*

- *both S and R are rational,*
- *each player knows which strategy the other player will play, and*
- *it is common knowledge between S and R that each of them is rational unless he follows the convention c.*

Then the strategy pair that is actually played is a fixed point of iterated best response, starting with c.

Conclusion

- rationality: standard assumption in Gricean pragmatics
- knowledge of the other player's startegy: precondition for sucessful communication ("meaning-nn")
- third condition bridges the gap between saying and meaning:
  - conventionalized semantics is a "(generalized) convention" in the technical sense
  - S and R pretend that they use the convention
  - if this leads to a uniqe fixed point under iterated best response, this fixed point describes what is pragmatically communicated

References

Grice, H. (1957). Meaning. *Philosophical Review*, **66**, 377–388.

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics 3: Speech Acts*, pages 41–58. Academic Press, New York.

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

Lewis, D. (1969). *Convention*. Harvard UP, Cambride (Mass.).

Stalnaker, R. (1997). On the evaluation of solution concepts. In M. O. L. Bacharach, L.-A. Gérard-Varet, P. Mongin, and H. S. Shin, editors, *Epistemic Logic and the Theory of Games and Decisions*, pages 345–64. Kluwer Academic Publisher.

Stalnaker, R. (2006). Saying and meaning, cheap talk and credibility. In A. Benz, G. Jäger, and R. van Rooij, editors, *Game Theory and Pragmatics*, pages 82–101. Palgrave MacMillan. to appear.