Explaining Quantity Implicatures

Tikitu de Jager*
Inst. for Logic, Language and Computation
Universiteit van Amsterdam
Amsterdam, The Netherlands
<S.T.deJager@uva.nl>

Robert van Rooij
Inst. for Logic, Language and Computation
Universiteit van Amsterdam
Amsterdam, The Netherlands
<R.A.M.vanRooij@uva.nl>

Abstract

We give derivations of two formal models of Gricean Quantity$_1$ implicature and strong exhaustivity (Van Rooij and Schulz, 2004; Schulz and Van Rooij, 2006), in bidirectional optimality theory and in a signalling games framework. We show that, under a unifying model based on signalling games, these interpretative strategies are game-theoretic equilibria when the speaker is known to be respectively minimally and maximally expert in the matter at hand. That is, in this framework the optimal strategy for communication depends on the degree of knowledge the speaker is known to have concerning the question she is answering.

In addition, and most importantly, we give a game-theoretic characterisation of the interpretation rule GRICE (formalising Quantity$_1$ implicature), showing that under natural conditions this interpretation rule occurs in the *unique* equilibrium play of the signalling game.

1 Introduction

An utterance in context is typically interpreted as having, in addition to its conventional, context-independent meaning, a CONVERSATIONAL IMPLICATURE that goes beyond the truth-conditional meaning. Particularly productive for analysing a large class of implicatures is the COOPERATIVE PRINCIPLE, introduced by Grice (1967): speakers may be assumed to try to contribute to the (jointly) accepted purpose of the conversation. We use a simplified version of Grice's maxim of Quality and focus on the first submaxim of Quantity, given as follows:

Definition 1 (Quality)**.** Say only what you know to be true.

Definition 2 (Quantity$_1$)**.** Make your contribution as informative as is required (for the current purposes of the exchange).

Given an utterance, the standard implicature via Quantity$_1$ is that the speaker did not intend to communicate any strictly stronger utterance (in the sense of truth-conditional entailment) from a contextually given set of alternatives.[1] For instance, if the question 'in the air' is how many children John has, the utterance "John has two children" standardly implicates that he does not have three or any greater number, i.e., that he has *exactly* two children.[2] However the strongest conclusion that can be drawn via Quantity$_1$ is that the speaker *does not know* that John has more children; the EXHAUSTIVE INTERPRETATION of the utterance says instead that she *knows that he does not* have more children. To reach this stronger interpretation various authors (see for example Spector, 2003; Van Rooij and Schulz, 2004) have recently suggested a two-stage approach: first the weak epistemic reading is derived by standard Gricean reasoning, then this is strengthened by the assumption that the speaker is an *expert*[3] in the matter at hand.

In this paper we examine a formal implementation of quantity implicature and exhaustive interpretation

*We would like to thank: Michael Franke for much helpful discussion, and for solving a technical problem with the derivation of QUANT in Bi-OT; Remko Scha for reminding us of the big picture; and Christopher Potts and the anonymous reviewers for comments regarding the manuscript.

[1] We will see in §2 that the choice of alternative expressions is crucial for even the simplest cases.

[2] That this is not truth-conditional meaning is easily seen: if the relevant question is *whether* he has (at least) two children —for tax purposes, say— then the utterance no longer carries the implicature and he might just as well have ten.

[3] Van Rooij and Schulz (2004) discuss speaker "competence", however this might give rise to confusion with the standard linguistic notion of the same name. We will therefore use the term "expertise" in this paper.

(originally proposed in the unpublished MA thesis of Katrin Schulz, and extended for exhaustification by Van Rooij and Schulz (2004); Schulz and Van Rooij (2006)), placed in the contexts of bidirectional optimality theory (Bi-OT) and of signalling games. We show firstly that given strong restrictions on the epistemic state of the speaker, quantity implicatures are *derivable* in Bi-OT. Next we show, under much weaker restrictions in the signalling games context, that (under a natural implementation of 'being expert') the interpretation according to quantity implicatures is rational just when the speaker is inexpert, and exhaustive interpretation just when she is expert.[4] We give, that is, a justification in terms of rational communication for the formalisation of $\mathsf{Quantity}_1$ and exhaustive interpretation given by Van Rooij and Schulz (2004); Schulz and Van Rooij (2006).

Finally, we show that in the game with an inexpert speaker, certain natural restrictions on the form of the interpretative strategy (convexity and faithfulness, defined in Theorem 29) completely characterise the strategy formalising quantity implicature: *only* that strategy can take part in a Nash equilibrium in the signalling game.

Notation

We write "$|\cdot|$" for the cardinality of a set. We assume throughout a set W of relevantly distinct possibilities, called for simplicity "worlds". Q is always a one-place predicate. If w is a world, then the valuation $V_w(Q)$ gives the set of objects satisfying Q in w.

We use an abbreviated notation for conditional probability: if $x \in X$ and $A \subseteq X$ then we abbreviate $P(\{x\} \mid A)$ by $P(x \mid A)$. (If P is a probability distribution over a set X, then for $A, B \subseteq X$ the conditional probability of A given B, $P(A \mid B)$, is standardly defined as $P(A \cap B)/P(B)$.)

The semantic denotation of an utterance f, $[\![f]\!]$, is the set of worlds in which f is true. We lift this standard notion to the level of information states (sets of worlds):[5]

$$(\!|f|\!) \stackrel{\text{def}}{=} \{s \subseteq [\![f]\!] \,;\, s \neq \varnothing\}.$$

Just as $[\![f]\!]$ gives the worlds in which f is true, $(\!|f|\!)$ gives the *information states* in which f is *licensed*, in the pragmatic sense of the maxim Quality. (Note that if s is an information state then $P(s \mid (\!|f|\!))$ is concise notation for $P(\{s\} \mid (\!|f|\!))$ and should not be confused with $P(s \mid [\![f]\!])$; the latter is only defined given a prior on individual worlds which we generally do not have in this setting.)

[4] 'Rationality' here is in the game-theoretic sense of playing a Nash equilibrium.

[5] The notation is due to Michael Franke, p.c.

Further notation will be introduced as needed.

2 $\mathsf{Quantity}_1$ and GRICE

Formalising $\mathsf{Quantity}_1$ requires establishing which utterances count as potential alternatives; if the utterance "John has *exactly* two children" is also an alternative, then this cannot be a $\mathsf{Quantity}_1$ implicature from "John has two children". The standard solution (taken by Horn (1972); Gazdar (1979); Levinson (2000), among others) is to consider only alternative expressions from a linearly ordered scale conventionally associated with the utterance (here the numerical expressions "John has (at least) n children"), hence the term SCALAR IMPLICATURE. The analysis given here generalises this approach to *partially* ordered alternative sets, the QUANTITY IMPLICATURE of utterances such as "John or Mary went to the party". The appropriate alternative expressions are the POSITIVE SENTENCES:

Definition 3 (Positive sentences)**.** A positive sentence contains only positive atoms, conjunction and disjunction. Given a finite domain of objects and a predicate Q, we define the finite set of POSITIVE Q-EXPRESSIONS by choosing a shortest exemplar from each class of logically equivalent positive sentences using only the predicate Q.

(The resulting denotations are all upwards monotonic in the question predicate; it is easy to see that introducing downwards-monotonic or non-monotonic expressions in general destroys the predictions. If "John and Mary *and nobody else*" is an alternative expression, then we will not strengthen the meaning of "John and Mary" via $\mathsf{Quantity}_1$, as desired.)

Suppose that the question 'in the air' is "Who went to the party?" and the answer given is "John or Mary went to the party". By quantity implicature we derive that the speaker does not know that both John and Mary attended. The stronger EXHAUSTIVE INTERPRETATION is that the speaker knows that John and Mary did not both attend the party.

In the framework described by Van Rooij and Schulz (2004); Schulz and Van Rooij (2006), scalar implicature is a special case of generalised quantity implicature, which is formalised as an interpretative principle "GRICE", incorporating both Quality and $\mathsf{Quantity}_1$. The 'Gricean interpretation' of an utterance f takes place against the background of a question predicate Q, the 'matter at hand', and the utterance is interpreted as meaning the set of states minimal with respect to speaker knowledge of Q ($\mathsf{Quantity}_1$) where f is known to hold (Quality).[6]

[6] We translate here the formal definitions of Van Rooij and

Definition 4 (Ordering by positive knowledge)**.** Let $s, s' \subseteq W$ be information states for the speaker. We say she HAS NO MORE POSITIVE KNOWLEDGE OF Q IN s THAN IN s', $s \leq^{\mathrm{K}}_{Q} s'$, (and analogously 'has *less*', $<^{\mathrm{K}}_{Q}$) when:

$$s \leq^{\mathrm{K}}_{Q} s' \stackrel{\text{def}}{\Longleftrightarrow} \forall w' \in s' : \exists w \in s : V_w(Q) \subseteq V_{w'}(Q).$$

(The ordering by positive knowledge only takes account of whether two information states differ in the positive Q-expressions they make true.)

Definition 5 (Minimising unstated positive knowledge)**.** The Gricean interpretation of an utterance f against a background predicate Q (according to Quality and $\mathsf{Quantity}_1$) is given by:

$$\begin{aligned}\textsc{Grice}(f, Q) &\stackrel{\text{def}}{=} \{s \in (\!|f|\!) \,;\, \forall s' \in (\!|f|\!) : s \leq^{\mathrm{K}}_{Q} s'\} \\ &= \{s \in (\!|f|\!) \;;\; \forall w' \in [\![f]\!] : \\ &\qquad \exists w \in s : V_w(Q) \subseteq V_{w'}(Q)\}.\end{aligned}$$

(That is, we restrict $\mathsf{Quantity}_1$ to informativity regarding *positive* knowledge; effectively this corresponds to the restriction to a language consisting only of positive expressions. Note also how the restriction to information states in $(\!|f|\!)$ corresponds to the requirement that Quality be fulfilled.)

As was already mentioned, this definition is not sufficient to derive the strong exhaustive readings standard for utterances like (in answer to "Who went to the party?") "John went to the party": not just that the speaker does not know of any other (relevant) individuals that they went, but that she knows that they did *not* go. The solution given by Van Rooij and Schulz (2004) (also Spector, 2003) is to assume the speaker is *maximally expert* with respect to Q (as far as this is consistent with the maxims).

Definition 6 (Ordering by expertise)**.** Let $s, s' \subseteq W$ be information states for the speaker. We say that she is NO MORE EXPERT ABOUT Q IN s THAN IN s', $s \leq^{\mathrm{E}}_{Q} s'$, (and analogously '*less* expert', $<^{\mathrm{E}}_{Q}$) when:

$$s \leq^{\mathrm{E}}_{Q} s' \stackrel{\text{def}}{\Longleftrightarrow} \forall w' \in s' : \exists w \in s : V_{w'}(Q) \subseteq V_w(Q).$$

Van Rooij and Schulz (2004) show that the 'strong epistemic' implicature of exhaustification can be correctly derived by maximising speaker expertise *after* applying GRICE:

Definition 7 (Maximising expertise)**.** Let f be a positive Q-expression. The exhaustive reading of f is given by

$$\begin{aligned}\textsc{Expert}(f, Q) \stackrel{\text{def}}{=} \{&s \in \textsc{Grice}(f, Q); \\ &\neg\exists s' \in \textsc{Grice}(f, Q) : s <^{\mathrm{E}}_{Q} s'\}.\end{aligned}$$

Next we show that GRICE, formalising both $\mathsf{Quantity}_1$ and Quality, can be derived from Quality alone in bidirectional optimality theory.

3 $\mathsf{Quantity}_1$ and Bi-OT

Bidirectional Optimality Theory (Bi-OT), first introduced by Blutner (2000), states that conventional language use is constrained by a *bidirectional* optimisation problem: a speaker should choose an optimal message to express her intent, and a hearer should choose the optimal interpretation for the message he hears. Optimality is expressed in terms of a relation $\succ$ ("is better than") between form/interpretation pairs, which we will return to in detail in a moment.

Definition 8 (Strong optimality)**.** Let $\langle f, s\rangle$ be a form/interpretation pair. We say $\langle f, s\rangle$ is STRONGLY OPTIMAL iff it satisfies the following two conditions:

$$\neg\exists s' : \langle f, s'\rangle \succ \langle f, s\rangle\,, \tag{1}$$
$$\neg\exists f' : \langle f', s\rangle \succ \langle f, s\rangle\,. \tag{2}$$

Following Blutner (2000), we define the ordering relation in terms of a complexity-related cost for forms, and the conditional probability of the interpretations given the lifted semantic meaning of the form (this second clause gives a preference for stereotypical interpretations):

Definition 9 (Bi-OT preference ordering)**.** Given forms f, f' and interpretations (information states) s, s', a probability distribution P on interpretations and a cost function "cost" mapping forms to numerical costs, the relation IS BETTER THAN, "$\succ$", is defined by

$$\begin{aligned}\langle f, s\rangle \succ \langle f, s'\rangle &\stackrel{\text{def}}{\Longleftrightarrow} P(s \,|\, (\!|f|\!)) > P(s' \,|\, (\!|f|\!)), \\ \langle f, s\rangle \succ \langle f', s\rangle &\stackrel{\text{def}}{\Longleftrightarrow} P(s \,|\, (\!|f|\!)) > P(s \,|\, (\!|f'|\!)) \\ &\qquad \text{or } P(s \,|\, (\!|f|\!)) = P(s \,|\, (\!|f'|\!)) \\ &\qquad\quad \&\ \mathrm{cost}(f) < \mathrm{cost}(f').\end{aligned}$$

We show now how, using strong optimality, we can derive the interpretative function GRICE given above.

3.1 Weak epistemic $\mathsf{Quantity}_1$ in Bi-OT

We will focus in this section on GRICE, the weak epistemic implicature ("... and *I don't know that* anyone

Schulz (2004) (which were given in a modal logic setting) to a formalism more amenable to the signalling games analysis in the sequel.

else came"). We will return to EXPERT in the signalling games setting (§4), where we focus on the difference between the weak epistemic interpretation and full exhaustification.

The following definition gives a probabilistic interpretation to "as informative as is required" in the maxim of $\mathsf{Quantity}_1$:

Definition 10 (Probabilistic quantity implicature)**.** Let W be a finite set of worlds differing in the extension of the predicate Q. Take F to be the full set of positive Q-expressions, and P a distribution on information states such that all interpretations are held possible.

Let $f \in F$ be an arbitrary utterance; the interpretations given by (probabilistic) QUANTITY IMPLICATURE from f are defined by

$$\textsc{Quant}(f) \stackrel{\text{def}}{=} \{s \in (\!|f|\!) \; ; \; \forall f' \in F : s \in (\!|f'|\!) \rightarrow P(s \,|\, (\!|f|\!)) \geq P(s \,|\, (\!|f'|\!))\}.$$

(That is, an information state s is in $\textsc{Quant}(f)$ if no alternative form makes s more likely than f does. Compare this to the definition of GRICE, which makes no explicit mention of the set of alternative forms at all.)

In the remainder of this section we show firstly that this definition corresponds (under a strong condition on the probability distribution over information states) to strong optimality, and finally that as an interpretative strategy it is equivalent to GRICE (thus retroactively justifying the name!).

Lemma 11. *If the probability distribution P over information states is* uniform *($P(s) = P(s')$ for all $s, s' \subseteq W$), and all forms are taken to have the same cost, then*

$$\textsc{Quant}(f) = \{s \subseteq W \; ; \; \langle f, s\rangle \text{ is strongly optimal}\}.$$

[Take $s \in \textsc{Quant}(f)$ arbitrary. Now $\langle f, s\rangle$ is not strongly optimal if there is a better $\langle f', s\rangle$ or $\langle f, s'\rangle$. Suppose the former; then for some $f' \neq f$ such that $s \in (\!|f'|\!)$, $P(s \,|\, (\!|f'|\!)) > P(s \,|\, (\!|f|\!))$ (since messages have equal cost) — but then $s \in \textsc{Quant}(f)$ is a contradiction. Suppose instead the latter; then for some $s' \neq s$ such that $s' \in (\!|f|\!)$, $P(s' \,|\, (\!|f|\!)) > P(s \,|\, (\!|f|\!))$, which is impossible since $s \in (\!|f|\!)$ and the distribution on information states is uniform.

For the converse, take $\langle f, s\rangle$ an arbitrary strongly optimal pair. Then $\neg\exists f' \in F : s \in (\!|f'|\!)$ & $P(s \,|\, (\!|f'|\!)) > P(s \,|\, (\!|f|\!))$; that is, $\forall f' \in F : s \in (\!|f'|\!) \rightarrow P(s \,|\, (\!|f|\!)) \geq P(s \,|\, (\!|f'|\!))$, thus $s \in \textsc{Quant}(f)$.]

This simple proof is included because it can easily be adopted to prove the simpler scalar implicature version of the above:

Corollary 12. *Take only messages arranged in a linear order by entailment (for instance the numerical expressions "John has n children" in the "at least" reading), and take* single *worlds as interpretations; as before, all messages are of equal cost and the distribution on interpretations is uniform. Then only the pairs $\langle$"John has n children"$, w\rangle$ where "John has* exactly *n children" is true in w are strongly optimal.*

(That is, precisely the standard —exhaustive— scalar implicature is produced by Bi-OT; adapting QUANT to probabilities of single worlds conditional on the *semantic* denotation gives the same result, by the proof given above. In fact, taking information states as interpretations leads to a weak epistemic version of scalar implicature, and maximising expertise in the sense of Definition 7 gives again the same result. We omit the details.)

Now the claim is that the interpretative principle QUANT formalises the maxims of $\mathsf{Quantity}_1$ and $\mathsf{Quality}$. Recall that the interpretative principle GRICE, Definition 5, is taken as a formalisation of precisely these maxims. We will now show that GRICE and QUANT are equivalent.

We define first a piece of helpful notation:

Definition 13 (Q-minimality)**.** Let s be a set of worlds and Q a question predicate. The Q-MINIMAL WORLDS in s are given by

$$\min_Q(s) \stackrel{\text{def}}{=} \{w \in s \; ; \; \neg\exists w' \in s : V_{w'}(Q) \subset V_w(Q)\}$$

Using $\min_Q(\cdot)$, for any information state s we can define a positive Q-expression f_s with the useful properties $s \subseteq [\![f_s]\!]$ and $\min_Q([\![f_s]\!]) = \min_Q(s)$ (that is, $[\![f_s]\!]$ is the upward completion of s in the ordering $\leq^{\mathrm{K}}_Q$). We omit the details.

Lemma 14. *Using $\min_Q(\cdot)$ we can formulate $\textsc{Grice}(f, Q)$ in two equivalent ways:*

$$\textsc{Grice}(f, Q) = \{s \in (\!|f|\!) \; ; \min_Q([\![f]\!]) \subseteq s\} \tag{3}$$

$$= \{s \in (\!|f|\!) \; ; \; \min_Q([\![f]\!]) = \min_Q(s)\}. \tag{4}$$

Now we are ready to show that QUANT and GRICE are equivalent.

Proposition 15. *Let Q be a one-place question predicate and f a positive Q-expression. Take all positive Q-expressions as alternatives to f, and information states (subsets of W) as interpretations. Assume that all information states have non-zero prior probability. Then*

$$\textsc{Grice}(f, Q) = \textsc{Quant}(f, Q).$$

Proof. We prove this by reducing $\textsc{Quant}(f, Q)$ to the reformulated definition of $\textsc{Grice}(f, Q)$ given in (3).

Lemma 16. *If s and s' are arbitrary sets of worlds and $\min_Q(s) \subseteq s' \subseteq s$ then $\min_Q(s) = \min_Q(s')$.*

[If w is Q-minimal in s, then no world in s' can give Q a smaller extension; if w is not Q-minimal in s then some $w' \in \min_Q(s)$ gives Q a smaller extension and $w' \in s'$.]

Lemma 17. *If f and f' are positive Q-expressions, then $[\![f]\!] = [\![f']\!] \iff \min_Q([\![f]\!]) = \min_Q([\![f']\!])$.*

[For any positive Q-expression f, an easy induction shows that $f_{[\![f]\!]} = f$. (In other words, positive Q-expressions are monotonic in Q.) The lemma follows immediately.]

Now for the reduction. First, take s and f such that $s \in (\!|f|\!)$ but $s \notin \textsc{Grice}(f, Q)$; by (3), then, $\min_Q([\![f]\!]) \not\subseteq s$. We will show that the form f_s (with denotation the upward completion of s) is preferable for s in Quant, just as it is in Grice.

Let s' be the (non-empty) information state $\min_Q([\![f]\!]) \setminus s$. Since $\min_Q(f_s) = \min_Q(s)$, $s' \not\subseteq [\![f_s]\!]$ so $[\![f_s]\!] \subset [\![f]\!]$ (by monotonicity of forms, $\min_Q([\![f_s]\!]) \subseteq [\![f]\!] \Rightarrow [\![f_s]\!] \subseteq [\![f]\!]$). But now by the assumption that no information state in $(\!|f|\!)$ is considered impossible, and since $s \subseteq [\![f_s]\!] \subset [\![f]\!]$ (so $s \in (\!|f_s|\!) \subset (\!|f|\!)$), we have $P(s \mid (\!|f|\!)) < P(s \mid (\!|f_s|\!))$, which implies $s \notin \textsc{Quant}(f, Q)$.

Next, take s such that $\min_Q([\![f]\!]) \subseteq s$, so $s \in \textsc{Grice}(f, Q)$, and suppose towards a contradiction that for some f' such that $s \in (\!|f'|\!)$ we have $P(s \mid (\!|f'|\!)) > P(s \mid (\!|f|\!))$ (so $s \notin \textsc{Quant}(f, Q)$). Since $s \in (\!|f|\!)$, for all alternative forms such that $[\![f']\!] \not\subseteq [\![f]\!]$ we have $P(s \mid (\!|f'|\!)) \leq P(s \mid (\!|f' \wedge f|\!))$ so without loss of generality we restrict ourselves to strengthenings of f.

Then by the probabilities $[\![f']\!] \neq [\![f]\!]$, so Lemma 17 gives us $\min_Q([\![f]\!]) \neq \min_Q([\![f']\!])$. But we have $\min_Q([\![f]\!]) \subseteq [\![f']\!] \subseteq [\![f]\!]$ which by Lemma 16 implies $\min_Q([\![f']\!]) = \min_Q([\![f]\!])$, a contradiction by Lemma 17. So no such f' can exist, which implies $s \in \textsc{Quant}(f, Q)$.

That is, $s \in \textsc{Quant}(f, Q) \iff \min_Q([\![f]\!]) \subseteq s \iff s \in \textsc{Grice}(f, Q)$. q.e.d

This means that under Bi-OT we can *derive* $\mathsf{Quantity}_1$ (represented by Grice) from Quality (required for optimality), but only under the assumption of a uniform distribution over information states. A Bayesian approach would suggest a *roughly* uniform prior in the absence of other information, but as long as we use Bi-OT as a predictive theory we require exact uniformity. In the following we will model an inexpert speaker as having a non-negligible probability of making inexact observations; the Bayesian ignorance view can be seen as a special case of this notion, but exact uniformity is clearly far too strong a requirement.

The important thing to note about Proposition 15 is that, unlike Lemma 11, the probability distribution need not necessarily be uniform, only everywhere non-zero. We show now, via the equivalence of Grice and Quant, that the same Gricean strategy is selected in a game-theoretic setting, under much more reasonable restrictions on the distribution.

4 $\mathsf{Quantity}_1$ and signalling games

Signalling games were introduced by Lewis (1969) to explain the existence of conventionalised meanings of linguistic expressions. In this context a typical model has multiple equilibria, and the choice of one among them is a matter of convention; Lewis's aim was indeed to show how such conventions might spontaneously arise without prior agreement (which would itself rely on a pre-existing conventional language). In this paper we use signalling games instead for *pragmatics*, in the tradition of Parikh (2001); Van Rooij (2004): we assume a predetermined (conventional) semantic meaning for the signals and show which (pragmatic) refinements of the semantic meaning are optimal in a game-theoretic sense. In contrast to the games of Lewis, here we have as desideratum a *single* equilibrium: given a pre-existing *semantic* convention, the *pragmatic* refinement should be uniquely determined. We will see that this goal is not reached in the basic model, but the addition in §5 of a structural constraint on the interpretative strategies gives us the refinement we are looking for.

Formally, a signalling game is a game of asymmetric incomplete information between a Sender and a Receiver, with chance moves by Nature. Given is a set of worlds W, a set of messages F, and a probability distribution P over information states (the moves of Nature). Nature shows an information state to the Sender, she sends a message f to the Receiver, and he plays an INTERPRETATION ACTION, a set of information states which we read as "the (perhaps pragmatically enriched) interpretation of f".[7]

Definition 18 (Interpretation signalling game)**.** An INTERPRETATION SIGNALLING GAME is a tuple $\langle Q, W, F, [\![\cdot]\!], \text{utility}_n, P\rangle$ with the following properties:

[7] Rather than define interpretation games in full generality, we include here the refinements specific to this application. A fully general definition would specify neither the objects taken as observations (and interpretations) nor the form of the utility function.

- Q, as above, is a one-place predicate;
- W is the full set of possibilities ("worlds") differing in the extension of Q;
- F is the set of positive Q-expressions;
- $[\![\cdot]\!]$ is the standard semantic denotation function (from which we derive also the lifted version $(\!|\cdot|\!)$);
- $\text{utility}_n \colon \wp(W) \times \wp(\wp(W)) \to \mathbb{R}$ is a function parameterised by n (as given below) which specifies the utility of each interpretation (by the receiver) in each information state (of the sender);
- P is a probability distribution on information states, such that all states in $\wp(W)$ occur with positive probability.

A PLAY of the game is a triple $\langle s, f, \mathcal{I}\rangle$ where $s \subseteq W$ is an information state (the OBSERVATION given by Nature to the Sender), $f \in F$ is a MESSAGE (sent by the Sender) and $\mathcal{I} \subseteq \wp(W)$ is a set of information states (the INTERPRETATION of the Receiver).

The UTILITY FUNCTION has the form

$$\text{utility}_n(s, \mathcal{I}) \stackrel{\text{def}}{=} \begin{cases} P(s \,|\, \mathcal{I}) & \text{if } s \in \mathcal{I}, \\ -n & \text{otherwise,} \end{cases}$$

where the parameter n is a large integer, the PENALTY for unsuccessful communication. This gives the PAYOFF of a play $\langle s, f, \mathcal{I}\rangle$ directly:[8]

$$\text{U}(s, f, \mathcal{I}) \stackrel{\text{def}}{=} \text{utility}_n(s, \mathcal{I}).$$

We will frequently wish to discuss families of games that vary only in their penalty values or observation distributions. Given a game $G = \langle Q, W, F, [\![\cdot]\!], \text{utility}_n, P\rangle$, a penalty value m and a distribution P', we write

$$\begin{aligned} G_m &\stackrel{\text{def}}{=} \langle Q, W, F, [\![\cdot]\!], \text{utility}_m, P\rangle\,, \\ G^{P'} &\stackrel{\text{def}}{=} \langle Q, W, F, [\![\cdot]\!], \text{utility}_n, P'\rangle\,, \text{ and} \\ G_m^{P'} &\stackrel{\text{def}}{=} \langle Q, W, F, [\![\cdot]\!], \text{utility}_m, P'\rangle\,. \end{aligned}$$

The penalty models the intuition that communicative failure is always the worst outcome, no matter how much effort is saved in arriving *efficiently* at a wrong interpretation. To see this, however, we need to be able to describe the strategies the two players use to produce their moves.

[8] In comparison to standard signalling games, this definition corresponds to the "cheap talk" assumption that messages are costless.

Definition 19 (Strategies)**.** A SENDER STRATEGY is a (total) function $\sigma \colon \wp(W) \to F$ giving a message in each information state. A RECEIVER STRATEGY is a (total) function $\rho \colon F \to \wp(\wp(W))$ giving an interpretation, a set of information states, for each message in F. A LANGUAGE is a pair $\langle \sigma, \rho\rangle$ where σ is a sender strategy and ρ is a receiver strategy.

A play $\langle s, f, \mathcal{I}\rangle$ is ACCORDING TO THE STRATEGIES σ and ρ if $\sigma(s) = f$ and $\rho(f) = \mathcal{I}$.

(The definitions of strategies for sender and receiver express the information asymmetry of the game. The sender observes (partially) the state of the world, but the receiver must use only the message in arriving at an interpretation.)

Now we can explain the necessity of the penalty value. Suppose this was absent (in the current model, simply set the penalty value n to zero). Now imagine that the same message, f, is sent in two information states s and s' that occur with equal (relatively high) probability. Then the interpretations $\rho(f) = \{s\}$, $\rho(f) = \{s'\}$ and $\rho(f) = \{s, s'\}$ all have equal payoff, so the third strategy, which never gives rise to communicative failure, is not preferred. Worse yet, if s is (even only very slightly) more probable than s', then this strategy is actually *worse* than $\rho(f) = \{s\}$: the strategy giving rise to communicative failure almost half the time is actually preferred. Setting a numerical value on the penalty for communicative failure turns out to be crucial for the analysis of expertise, as we will see in §4.1.

Definition 20 (Payoffs)**.** Given a game $G = \langle Q, W, F, [\![\cdot]\!], \text{utility}, P\rangle$ and a language $L = \langle \sigma, \rho\rangle$ for G, the EXPECTED UTILITY (or PAYOFF) of L in G, $\text{EU}_G(\sigma, \rho)$, is given by

$$\sum \{P(s) \cdot \text{U}(s, f, \mathcal{I}); \quad \langle s, f, \mathcal{I}\rangle \text{ is a play according to } \sigma \text{ and } \rho\}.$$

A sender strategy σ is a BEST SENDER RESPONSE to a receiver strategy ρ in G if for all $\sigma' \neq \sigma$, $\text{EU}_G(\sigma, \rho) \geq \text{EU}_G(\sigma', \rho)$, a STRICT BEST RESPONSE if the inequality is everywhere strict, and analogously for best receiver responses.

We define also the expected utility of σ and ρ AT AN INFORMATION STATE s (or AT A MESSAGE f) by taking expectations over only those plays according to σ and ρ that include the information state (or message), and write this $\text{EU}(\sigma, \rho)(s)$ (or $\text{EU}(\sigma, \rho)(f)$).

We use the standard game-theoretic notion of Nash equilibrium to single out certain preferred strategies: a pair of rational agents will play a language that is a Nash equilibrium because if the language is *not* an equilibrium, then some player has a payoff incentive to change their strategy.

Definition 21 (Nash equilibrium)**.** A language $\langle \sigma, \rho \rangle$ is a NASH EQUILIBRIUM for the game G if σ is a best reply to ρ and vice versa. It is a STRICT NASH EQUILIBRIUM if σ and ρ are mutual *strict* (i.e., unique) best replies.

Now we give two related games, representing respectively an inexpert and an expert speaker, and show that the Nash equilibrium solution concept selects the strategies we want in each case.

4.1 Expertise in signalling games

The translation of the notion of 'expertise' to the signalling game proceeds via the probability distribution on observations (with appropriate adjustments of the penalty parameter in the utility function). The intuition is that an expert speaker is more likely to make precise observations, whereas with an inexpert speaker we cannot know whether their utterance was prompted by exact knowledge of the situation or by an extremely vague observation. So for a maximally inexpert speaker we expect all information states to appear with non-negligible probability.

As was stated earlier, we model Quality via the lifted semantic denotation function taking messages to the information states that license them. That is, we consider only sender strategies σ that satisfy, for all information states s, $s \in (\!|\sigma(s)|\!)$. We will take GRICE and EXPERT as *interpretative* strategies, defining receiver strategies ρ_{G} and ρ_{E}:

$$\rho_{\mathrm{G}}(f) \stackrel{\mathrm{def}}{=} \textsc{Grice}(f, Q), \qquad \text{(Cf. Definition 5)}$$
$$\rho_{\mathrm{E}}(f) \stackrel{\mathrm{def}}{=} \textsc{Expert}(f, Q). \qquad \text{(Cf. Definition 7)}$$

Definition 22 (Inexpert speaker)**.** Take $\delta \in [0, 1]$ a non-negligible value. The distribution P represents an INEXPERT SPEAKER (with respect to δ) if $\forall s \subseteq W : P(s) \geq \delta$.[9]

Proposition 23. *Let P_δ represent an inexpert speaker for some given value of δ. Then by choice of n we can always construct a game $G_n^{P_\delta}$ with the following property:*

For any sender strategy σ utilising all messages, the unique *best receiver reply $\rho_{\mathrm{BR}(\sigma)}$ is given by*

$$\rho_{\mathrm{BR}(\sigma)}(f) = \sigma^{-1}(f) \stackrel{def}{=} \{s \subseteq W \,;\, \sigma(s) = f\}.$$

We call such a game a GAME WITH INEXPERT SPEAKER.

Theorem 24. *Let G be a game with inexpert speaker. Then in G,*

1. *ρ_{G} has a unique best sender reply σ_{G}, and $\langle \sigma_{\mathrm{G}}, \rho_{\mathrm{G}} \rangle$ is a strict Nash equilibrium; and*
2. *for no sender strategy σ is $\langle \sigma, \rho_{\mathrm{E}} \rangle$ a Nash equilibrium.*

Proof of clause 1. Let ρ be an arbitrary receiver strategy for the game $G_{\epsilon,n}$. Then the speaker's best reply $\sigma_{\mathrm{BR}(\rho)}$ is given by

$$\sigma_{\mathrm{BR}(\rho)}(s) = \arg\max_{f \in F} \{P(s \,|\, \rho(f)) \,;\, s \in \rho(f)\}.$$

For ρ_{G}, the receiver strategy assigning to each message f its Gricean interpretation $\textsc{Grice}(f, Q)$, this is

$$\sigma_{\mathrm{BR(G)}}(s) = \arg\max_{f \in F} \{P(s \,|\, \rho(f)) \,;\, s \in \textsc{Grice}(f, Q)\}.$$

But now a property of GRICE makes this apparent optimisation problem trivial: each information state s occurs in $\textsc{Grice}(f, Q)$ for *exactly one message* f. The set $\{f \,;\, s \in \textsc{Grice}(f, Q)\}$ is a singleton. So the "arg max" is redundant, the best message to use for s is the *only* message to use for s.

To see this, recall that the form f_s is the minimal (in terms of set inclusion) message whose denotation contains s: $s \subseteq [\![f]\!] \Rightarrow \min_Q(s) \subseteq [\![f]\!] \Rightarrow [\![f_s]\!] \subseteq [\![f]\!]$, since $\min_Q(s) = \min_Q([\![f_s]\!])$ generates $[\![f_s]\!]$ by the Q-monotonicity condition on forms. Now if we assume that all information states are held possible, $s \subseteq [\![f_s]\!] \subset [\![f]\!] \Rightarrow s \in (\!|f_s|\!) \subset (\!|f|\!) \Rightarrow P(s \,|\, (\!|f_s|\!)) > P(s \,|\, (\!|f|\!))$, so $s \notin \textsc{Grice}(f, Q)$. That is, under these assumptions, each information state gives rise to a unique optimal message.[10]

If we take this observation to define GRICE as a strategy for production, $\sigma_{\mathrm{G}}(s) \stackrel{\mathrm{def}}{=} f_s$, then it is easy to see that $\textsc{Grice} = \langle \sigma_{\mathrm{G}}, \rho_{\mathrm{G}} \rangle$ is its own best response:

$$\begin{aligned} \rho_{\mathrm{BR}(\sigma)}(f) &= \{s \,;\, \sigma(s) = f\} \\ \rho_{\mathrm{BR(G)}}(f) &= \{s \,;\, \sigma_{\mathrm{G}}(s) = f\} \\ &= \{s \,;\, f_s = f\} \\ &= \{s \,;\, \min_Q(s) = \min_Q([\![f]\!])\} \\ &= \rho_{\mathrm{G}}(f) \end{aligned}$$

(by the rephrased definition of $\textsc{Grice}(f, Q)$, (4) from Lemma 14).

That is, GRICE interpreted in this way as strategies $\langle \sigma_{\mathrm{G}}, \rho_{\mathrm{G}} \rangle$ is its own unique best reply, a strict Nash equilibrium. q.e.d

[9] Clearly for values of δ larger than the reciprocal of the number of information states no distribution will satisfy the condition.

[10] Compare for instance mention-some questions ("Where can I buy a newspaper?"), in which for many information states several answers are intuitively equally optimal; these require a different payoff function, and will not be treated further in this paper.

The second clause of Theorem 24 is easy to see. Let σ be a strategy for which ρ_{E} is a putative best reply; since σ is total, a best receiver response $\rho_{\mathrm{BR}(\sigma)}$ (as given by Proposition 23) should include every information state in the interpretation of some message. But ρ_{E} does not do this (some information states are UNINDUCABLE), so the payoff according to ρ_{E} falls short of that given by $\rho_{\mathrm{BR}(\sigma)}$ according to the penalty.

Definition 25 (Expert speaker)**.** Take $\epsilon \in [0,1]$ a value 'reasonably close' to zero. The distribution P represents an EXPERT SPEAKER (with respect to ϵ) if for all $s \subseteq W$, $P(s) < \epsilon$ just in case $\exists w, w' \in s : V_w(Q) \subset V_{w'}(Q)$.

(Note that this definition is a rough parallel to Definition 6, in that the observations receiving low probability are each less expert than some other observation that receives high probability.)

Proposition 26. *Let G_n be a game with penalty n. We can find a positive probability ϵ with the following property:*

For all *(everywhere-nonzero) probability distributions P, in G_n^P the following holds for each receiver strategy ρ: If for any $s, s' \subseteq W$ and $f \in F$ we have $\{s, s'\} \subseteq \rho(f)$ and $P(s) < \epsilon \leq P(s')$, then ρ is not the best response to any sender strategy in G_n^P.*

A game G_n^P is known as a GAME WITH EXPERT SPEAKER if there is such an ϵ for G_n such that P represents an expert speaker with respect to ϵ.

Intuitively the triple $\langle \delta, n, \epsilon \rangle$ as a whole represents a 'cultural parameter' of language use, corresponding roughly to the notion of how much evidence is 'enough' to justify stating something with conviction. It has been suggested, on the basis of data from Malagasy, that quantity implicature is not in fact universal (Keenan, 1977); an alternative interpretation of the data seems to be that the 'required degree of conviction' parameter is in this case turned extremely high. It is interesting to speculate whether this notion could be adapted for representing evidential markers (see for example Ifantidou, 2001).

Lemma 27. *If G_n and G'_n are games with respectively inexpert and expert speakers parameterised by δ and ϵ but sharing the penalty value n, then $\epsilon \leq \delta$.*

(In other words, the notions of inexpertise and expertise in use here are compatible, but represent in general extremes rather than a binary opposition.)

Theorem 28. *Let G be a game with expert speaker. Then in G,*

1. *for no sender strategy σ is ρ_{G} a best response; and*
2. *there is at least one σ which is a best sender response to ρ_{E}, and for* every *such σ, ρ_{E} is in turn the unique best receiver response.*[11]

The first clause follows immediately from the construction of the game with expert speaker. The second clause is a generalisation of the notion of Nash equilibrium, which is necessary in the signalling games setting when (as in this case) some information states are uninducable.[12] We cannot (as in Theorem 24) simply find a sender strategy forming a strict Nash equilibrium: how such a strategy behaves on the uninducable —low probability— information states will not affect the payoff, so there will be many non-strict best responses. What is ensured by the construction, however, is that all of these sender strategies will behave the same way on the high-probability information states, and thus that ρ_{G} will be the unique best receiver response to each of them.

5 GRICE characterised

In the previous section we showed that playing according to GRICE is rational when the speaker is inexpert, and according to EXPERT when she is expert. However the question still remains, what of other possible strategies? In the case of an inexpert speaker, the game also admits of a multitude of alternative solutions, some decidedly pathological-looking. We would like to do more than show that GRICE is rational, we would like to show that it is the *only* rational strategy given an inexpert speaker. The following characterisation result comes much closer to achieving this desideratum:

Theorem 29. *Let G be a game with inexpert speaker. Let ρ be a receiver strategy with the following properties:*

1. *$\forall f \in F : [\![f]\!] \in \rho(f)$ ("Faithfulness"), and*
2. *$\forall s, s', s'' \subseteq W : \forall f \in F : s \subseteq s' \subseteq s''$ & $s, s'' \in \rho(f) \Rightarrow s' \in \rho(f)$ ("Convexity").*

Then there exists a sender strategy σ (obeying Quality*) such that $\langle \sigma, \rho \rangle$ is a strict Nash equilibrium in G if and only if $\rho(f) = \textsc{Grice}(f, Q)$.*

[11] The formulation is a special case of the notion of an EVOLUTIONARILY STABLE SET of languages: it is —roughly speaking— a maximal closed set of neutrally stable (mixed) strategies surrounded by a region of strategies that earn lower payoff, in this case with the additional property that each language appearing in the set has the same sender strategy.

[12] In more general terms: when there are more states than messages, and when mixed strategies are disallowed.

Clearly these conditions alone are not sufficient to characterise GRICE. The requirement of rational play in a game with inexpert speaker ensures that $\min_Q([\![f]\!]) \in \rho(f)$ for each message f; this provides a minimal element for each interpretation. Faithfulness provides a maximal element, and Convexity fills in the information states between to match GRICE.

Proof. First, let $\rho(f) = \textsc{Grice}(f, Q)$ for all $f \in F$. By Lemma 14 (4), $[\![f]\!] \in \textsc{Grice}(f, Q)$ for all f, so Faithfulness is fulfilled. Now take $s, s', s'' \subseteq W$ and $f \in F$ such that $s, s'' \in \textsc{Grice}(f, Q)$. Then by Lemma 14 (3),

$$\begin{aligned} &\min_Q([\![f]\!]) \subseteq s \subseteq s' \subseteq s'' \subseteq [\![f]\!] \\ &\Rightarrow \min_Q([\![f]\!]) \subseteq s' \subseteq [\![f]\!] \\ &\Rightarrow s' \in \textsc{Grice}(f, Q), \end{aligned}$$

so the Convexity condition is also satisfied.

The converse is a little more involved. Suppose that ρ is both faithful and convex (in the sense of Theorem 29). It is sufficient to show that if ρ occurs in a Nash equilibrium language, then for all f, $\min_Q([\![f]\!]) \in \rho(f)$. [In that case by faithfulness $[\![f]\!] \in \rho(f)$ and by the convexity condition all $s \subseteq [\![f]\!]$ such that $\min_Q(f) \subseteq s \subseteq [\![f]\!]$ also occur in $\rho(f)$; that is, $\rho(f) \supseteq \textsc{Grice}(f, Q)$. If $\rho(f) \neq \textsc{Grice}(f, Q)$, then some information state s' outside $\textsc{Grice}(f, Q)$ is also in $\rho(f)$; then s' will also appear in $\rho(f')$ for some $f' \neq f$, so ρ is not the best response to any sender strategy, a contradiction.]

Let ρ occur in a Nash equilibrium language. Suppose towards a contradiction that for some $f' \neq f$ that $\min_Q([\![f]\!]) \in \rho(f')$. Let σ be a best response to ρ; then $\sigma(\min_Q([\![f]\!])) = f'$. If σ obeys Quality, since message denotations are upwards monotonic we have $\min_Q([\![f]\!]) \subseteq [\![f]\!] \subseteq [\![f']\!]$. But since $\{\min_Q([\![f]\!]), [\![f']\!]\} \subseteq \rho(f')$ (by faithfulness and our hypothesis), by the convexity condition we have also $[\![f]\!] \in \rho(f')$. But now by faithfulness we have also $[\![f]\!] \in \rho(f)$, and this means, by Proposition 23, that ρ is not a best reply to σ, or indeed to *any* sender strategy.

That is, if ρ satisfies the conditions of Theorem 29 and occurs in a Nash equilibrium language, then for all f, $\min_Q([\![f]\!])$ occurs only in $\rho(f)$; this in turn implies that $\rho(f) = \textsc{Grice}(f, Q)$, and the equivalence is complete. q.e.d

The names of the conditions in Theorem 29 are deliberately suggestive. These are not arbitrary properties, they are very natural constraints on the structure of an interpretative principle.

The first, Faithfulness, is perhaps even stronger: it can be read as a necessary condition on the relation between semantics and pragmatics in the model. Recall that we began by simply stipulating a conventionalised semantic meaning for each of our messages. If the 'semantic meaning' of some message is *never* in the interpretation by a receiver, it could be described as somewhat disingenuous to continue to call it 'semantic meaning'; the faithfulness requirement then ensures that our terminology remains honest.

The convexity condition is a closure property on sets; like any such, it helps enormously for describing, learning, and remembering the interpretations these sets represent, since we can compactly describe a set in terms of just a few of its elements. In our case, we can describe the pragmatic interpretation of f in terms of just $\min_Q([\![f]\!])$, once the two constraints are given. Convexity constraints in particular play an important role in describing linguistic universals in generalised quantifier theory (Thijsse, 1983; Van Benthem, 1986) and cognitive semantics (Gärdenfors, 2000), and are given an independent game-theoretic motivation in a forthcoming paper by Jäger and Van Rooij.

6 Conclusion

We have given a game-theoretic implementation of the interpretative principles GRICE (Quantity$_1$ implicature) and EXPERT (exhaustive interpretation) defined by Van Rooij and Schulz (2004); Schulz and Van Rooij (2006). In a signalling games framework we specified what it means to have an expert speaker by means of the penalty value and 'degree of conviction' parameter bundle $\langle \delta, n, \epsilon \rangle$. Under these definitions, we found that *interpretation* according to GRICE and EXPERT is rational in exactly the cases we would expect; GRICE induces a strict Nash equilibrium in the inexpert case and thus fixes the sender strategy, while EXPERT in the expert case leaves some details of the sender strategy unspecified but is nonetheless stable in a natural extended sense.

These models did not achieve the desideratum of singling out a unique pragmatic interpretation rule. To do this we require in addition *structural* constraints on the form of an interpretative strategy: Faithfulness (that pragmatic interpretation should not discard the semantic meaning) and Convexity (that the interpretation of a message should be a convex set, under the set inclusion partial order). With these constraints, and with an inexpert speaker, the uniquely rational interpretative strategy is Gricean Quantity$_1$ implicature as modeled by GRICE.

Note also that all these results have an interpretation in *evolutionary* game theory: strict Nash equi-

librium corresponds to evolutionary stability, while the extended equilibrium notion required for the expert speaker is closely related to neutral stability and the notion of an evolutionarily stable *set* of strategies. In particular, the characterisation result implies that GRICE is the unique evolutionarily stable strategy for the game with inexpert speaker.

This perspective suggests a different way of looking at the relationship between GRICE and EXPERT. If we consider *speaker* strategies instead of interpretative strategies, Theorem 28 shows that EXPERT is nothing but GRICE restricted to the information states of an expert speaker. Combined with the characterisation result of Theorem 29 we are left with a picture of GRICE as a fundamental pragmatic rule and EXPERT as an application of that rule in the common case of a speaker we trust to know what she is talking about.

References

Johan F. A. K. van Benthem. *Essays in Logical Semantics*. Reidel, Dordrecht, 1986.

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

Peter Gärdenfors. *Conceptual Spaces: The Geometry of Thought*. MIT Press, Cambridge, Massachusetts, 2000.

Gerald Gazdar. *Pragmatics*. Academic Press, London, 1979.

H. P. Grice. Logic and conversation. The William James Lectures, delivered at Harvard University. Republished with revisions in Grice (1989), 1967.

H. P. Grice. *Studies in the Way of Words*. Harvard University Press, Cambridge, Massachusetts, 1989.

Laurence R. Horn. *The semantics of logical operators in English*. PhD thesis, Yale University, 1972.

Elly Ifantidou. *Evidentials and Relevance*, volume 86 of *Pragmatics & Beyond New Series*. John Benjamins, 2001.

Gerhard Jäger and Robert van Rooij. Language structure: psychological and structural constraints. *Synthese*, to appear.

Elinor Keenan [Ochs]. On the universality of conversational implicatures. In Ralph W. Fasold and Roger W. Shuy, editors, *Studies in Language Variation: semantics, syntax, phonology, pragmatics, social situations, ethnographic approaches*, pages 255–269, Washington, D.C., 1977. Georgetown University Press.

Stephen C. Levinson. *Presumptive Meanings. The Theory of Generalized Conversational Implicatures*. MIT Press, Cambridge, Massachusetts, 2000.

David K. Lewis. *Convention*. Harvard University Press, Cambridge, 1969.

Prashant Parikh. *The Use of Language*. CSLI Publications, Stanford, California, 2001.

Robert van Rooij. Signalling games select Horn strategies. *Linguistics and Philosophy*, 2004.

Robert van Rooij and Katrin Schulz. Exhaustive interpretation of complex sentences. *Journal of Logic, Language and Information*, 13:491–519, 2004.

Katrin Schulz and Robert van Rooij. Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy*, 29:205–250, 2006.

Benjamin Spector. Scalar implicatures: exhaustivity and Gricean reasoning? In Balder ten Cate, editor, *Proceedings of the Eighth ESSLLI Student Session*, Vienna, Austria, August 2003.

Elias Thijsse. On some proposed universals of natural language. In Alice G. B. ter Meulen, editor, *Studies in Modeltheoretic Semantics*, pages 19–36. Foris Publications, Dordrecht, 1983.