Quantity Implicatures, Exhaustive Interpretation, and Rational Conversation

Michael Franke
Department of Linguistics
Universtity of Tübingen

October 30, 2010

Abstract

Quantity implicatures are inferences triggered by an utterance based on what other utterances a speaker could have made instead. Using ideas and formalisms from game theory, I demonstrate that these inferences can be explained in a strictly Gricean sense as *rational behavior*. To this end, I offer a procedure for constructing the context of utterance insofar as it is relevant for quantity reasoning as a game between speaker and hearer. I then give a new solution concept that improves on classical equilibrium approaches in that it uniquely selects the desired "pragmatically correct" play in these interpretation games by a chain of back-and-forth reasoning about the opponent's behavior. To make this formal approach more accessible to a wider audience, I give a simple algorithm with the help of which the model's solution can be computed without having to do heavy calculations of probabilities, expected utilities and the like. Subsequently, I show that this rationalistic approach subsumes and improves on recent exhaustivity-based approaches in that the former predicts uniformly for quantity implicatures of various epistemic varieties, free-choice readings of disjunctions, as well as a phenomenon tightly related to the latter, namely so-called "simplification of disjunctive antecedents".

1 Introduction

In his essay *Logic & Conversation* (Grice 1975), Paul Grice made a beautiful case for parsimony of a theory of meaning in a defense of logical semantics. Grice maintained that semantic extravagance is often unnecessary, and he showed that many alleged differences between classical logical semantics

and intuition can be explained systematically as arising from certain regularities of our conversational practices, which he summarized under the label *Maxims of Conversation*. Based on the assumption that the speaker adheres to these maxims, the hearer is able to pragmatically enrich the semantic meaning of an utterance in systematic ways. This is, in fairly simplified terms, the background of Grice's theory of *conversational implicatures*.[1]

A particular kind of conversational implicature are so-called *quantity implicatures*. For illustration, consider the following conversation:

(1) JACK: I often start my yoga practice with headstand.
JILL: I sometimes do that too.

From Jill's reply in (1), we readily and reasonably infer the following:

(2) It's not the case that Jill starts her yoga practice with headstand often.

Does that mean that the temporal adverbial *sometimes* means "sometimes but not often"? Absolutely not, Grice would say. The literal meaning of *sometimes* may very well be just an existential quantification over times or events. This is because we can explain the inference in (2) as something that Jack, the hearer, is entitled to infer based on the assumption that Jill, the speaker, is forthcoming and cooperative towards the interest shared by speaker and hearer, which is —so the central Gricean assumption goes— honest and reliable information exchange. In other words, Jack may reasonably assume that Jill's linguistic behavior is governed by the following *Speaker Quantity Principle*:[2]

(3) SPEAKER QUANTITY PRINCIPLE:
A cooperative speaker provides all true and relevant information she is capable of.

Based on this, Jack may reason that it would have been more informative —as far as logical strength goes— for Jill to simply have responded "Me too." The extra information would have been relevant, at least for the sake of small-talk conversation. Hence, the reason why Jill has not simply said "Me too" was most likely so as not to speak untruthfully. So it must be the case that (2) is true.

[1] Grice's theory of conversational implicatures was developed further in many ways by a great number of linguists and philosophers. I do not wish to claim that I am producing Grice's original view on the matter, but rather that which I take to be the distilled common idea that emerged and is still developed to-day in the community. Crucial steps in the shaping of our ideas about conversational implicatures were made by, among others, Horn (1972), Gazdar (1979), Atlas and Levinson (1981), Levinson (1983), Horn (1984), Horn (1989), Levinson (2000).

[2] This principle is not found in Grice's writings, but is a simplified and condensed conglomerate of Grice's Maxims of Quantity, Quality and Relevance. The present formulation does echo, however, rather closely the formulation given by Schulz and van Rooij (2006) in motivation of their exhaustifity operator (to be introduced in Section 4.1).

This sketchy derivation clearly needs to be improved if it is to be the backbone of a respectable theory of quantity implicature *tout-court*. Firstly, it is clearly desirable to trade in the above informal reasoning for a more perspicuous, calculable and formal theory of quantity inferences, especially when we wish to assess predictions in more involved cases. This has been done, with ample degree of success (e.g. Gazdar 1979; Schulz and van Rooij 2006; Spector 2006). But it is also fair to say that the more formally rigorous a theory of quantity implicature is, the further removed it usually stands from the original idea behind the Gricean approach to conversational implicatures (cf. Geurts and Pouscoulous 2009*b*, pp. 1–3). Which is what exactly?

Grice was an extremely careful philosopher, and as such he was naturally very aware of the tentativeness of his own proposal (cf. Chapman 2005). Indeed, Grice envisaged a *rationalistic foundation* of the Maxims of Conversation and the explanations of conversational implicatures that they license:

> "As one of my avowed aims is to see talking as a special case or variety of purposive, indeed rational, behaviour, it may be worth noting that the specific expectations or presumptions connected with at least some of the foregoing maxims have their analogues in the sphere of transactions that are not talk exchanges."
>
> (Grice 1989, p. 28)

This essay is in this general spirit of Grice's programme. The main question to be asked is whether complex cases of quantity implicatures, can be grounded in rationalistic terms, using the formal machinery of game theory. After recapitulating some of the relevant issues concerning the notion of quantity implicature in Section 2, we will review a test set of quantity implicatures in Section 3, some of which have proven problematic even for our currently best formal theories, which I take to be theories based on the idea of *exhaustive interpretation*. This is followed by a brief assessment of these theories in Section 4, assessing some of their weaknesses and strengths. Section 5 then addresses the general question what a rationalistic explanation of quantity implicatures would have to look like. This paves the way for the game-theoretic model to be introduced in Sections 6, 7 and 8. Section 9 discusses the model's predictions, before Section 10 concludes.

2 Quantity Implicatures: Background

Epistemic Variety. The derivation of a quantity implicature crucially involves comparing the utterance the speaker made with *alternative utterances* the speaker could have made. For instance, the utterance of

(4) I sometimes practice headstand.

would have to be compared to, at least, the logically stronger alternative:

(5) I often practice headstand.

This comparison may give rise to any of the following varieties of a quantity implicature: the *general epistemic implicature* in (6a), the *strong epistemic implicature* in (6b), the *weak epistemic implicature* in (6c), or the *factual quantity implicature* in (6d).[3]

(6) a. The speaker does not believe that she practices headstand often.
b. The speaker believes that she does not practice headstand often.
c. The speaker is uncertain whether she practices headstand often.
d. The speaker does not practice headstand often.

The general epistemic implicature in (6a) should be considered most basic, because it is this inference only that we are strictly speaking entitled to draw when a principle like (3) tells us that the speaker was not able to truthfully assert the stronger statement (5). The strong epistemic implicature (6b) and the factual implicature (6d) can then in fact be derived from the general epistemic implicature under an additional assumption of *sender competence*:[4] if we assume that the speaker knows how frequently she practices headstand, then (6a) may be strengthened to (6b), which may in turn be strengthened to (6d). I will come back to define a formal rendering of competence in Section 6.2. For the time being, the reader's intuitive understanding suffices entirely.

It is often overlooked that the general epistemic implicature in (6a) can also be strengthened the other way (see Geurts 2010, Chapter 2.3): if we assume that the speaker is *in*competent, i.e., that she is likely *not* informed properly, then we can strengthen (6a) to yield the weak epistemic implicature in (6c) too.

The general logic of belief I adopt here is the standard possible-worlds analysis of Hintikka. If φ is a proposition (a set of possible worlds), then $\mathrm{Bel}_S\varphi$ expresses that the speaker believes that φ is true. (This is the case, as is classically assumed, just when all the possible worlds which the speaker considers possible are contained in φ.) I use notation $\mathrm{Uc}_S\varphi$ to denote that the speaker is uncertain whether φ is true, which is an abbreviation of $\neg\mathrm{Bel}_S\varphi \wedge \neg\mathrm{Bel}_S\neg\varphi$.

[3]Notice that, for various reasons, I spell out epistemic implicatures in terms of speaker beliefs, as do, for instance, Sauerland (2004) or Geurts (2010). Others prefer the stronger formulation in terms of knowledge (e.g. Gazdar 1979; Soames 1982; Schulz and van Rooij 2006; Russell 2006).

[4]There are several conceptually and formally different implementations of this idea in the literature (van Rooij and Schulz 2004; Sauerland 2004; Schulz 2005; Schulz and van Rooij 2006; Russell 2006; Geurts 2010).

With this there are then three *belief values*, as I will call them, three possibilities for an agent's beliefs to relate to a proposition: (i) the agent believes the proposition is true, (ii) believes that it is false, or (iii) is uncertain about it.

Alternatives: Unde Venitis? Which quantity implicatures can be derived obviously hinges on which expression alternatives we consider. If we had considered also the logically stronger expression:

(7) I sometimes but not often practice headstand.

derivation of the factual implicature in (6d) would have been jeopardized.[5] This is why, traditionally, *expression alternatives* were taken to be derived by lexical substitution of elements that are associated on an ordered *scale* of alternatives (e.g. Horn 1972, 1984).[6] To exclude pathological predictions, these scales were required to satisfy certain conditions, such as that elements had to be equally lexicalized, should not be more complex than the original utterance, or had to have the same monotonicity properties (e.g. Atlas and Levinson 1981; Horn 1989; Matsumoto 1995). Some authors have since generalized the notion of an ordered scale to that of an unordered *set* of lexical alternatives (e.g. Sauerland 2004; Fox 2007), others have been even more liberal, or so it seems, by comparing utterances to the set of possible answers to some possibly implicit question under discussion (van Kuppevelt 1996; van Rooij and Schulz 2004; Schulz and van Rooij 2006; Spector 2006). The issue which alternatives to consider for derivation of quantity implicatures is, however, still ongoing (cf. Katzir 2007; Swanson 2010).

I have no new theory to offer on this question either. The system I am going to introduce in the following sections actually is not a theory of alternatives, but one of reasoning *about* alternatives: given a set of alternatives as input, it will specify which implicatures we may expect. Still, the game theoretical perspective that I am going to defend in the following does indirectly add to the debate. By explicitly modeling a speaker and a hearer and their beliefs about each other's beliefs and action alternatives, we are reminded of the fact that a conversational implicature is not something that attaches to a sentence in isolation. A conversational implicature, in the original Gricean sense, can only be derived from an utterance (cf. Bach 2006; Geurts 2010). If this is right, then the hearer's choice which alternatives to consider must be

[5] It is still possible to derive an epistemic implicature from an utterance of (4) if we consider both of the alternative utterances (5) and (7). In that case, we obtain that the speaker is uncertain whether (5) or (7) is true. This partly speaks to the ominous "symmetry problem" (e.g. Block 2008; Katzir 2007)

[6] Quantity implicatures derived with the help of such a scale are therefore often referred to as *scalar implicatures*. Here, I negligently use this term for the special case of quantity implicature associated with the lexical contrast between "some" and "all" (see Section 3).

made in context, based on assumptions about which alternatives the speaker may have been aware of in the first place, and which of those she considered worthwhile entertaining.[7] But that means that factors such as strength of lexical association, morphosyntactic complexity, and lexical semantic properties may all play *a* role in the hearer's construction of his assessment of the context of utterance, but it is unlikely, to say the least, that any *one* factor should be solely decisive (see Swanson (2010) for a similar opinion). In the following I will therefore largely rely on an intuitive understanding of what may count as a consistent, yet *prima facie* plausible set of alternative expressions in a given context.

3 Quantity Implicatures: Some Relevant Cases

This paper will be mainly concerned with the interpretation of three types of disjunctive constructions: (i) *plain disjunctions* of the form $A \vee B$, (ii) *free-choice disjunctions* of the form $\Diamond(A \vee B)$ and (iii) disjunctions in antecedents of conditionals such as $(A \vee B) > C$. This sections summarizes what is worth explaining about these constructions. As an additional test case, we should also consider simple "scalar implicatures".

Scalar Implicature. To start with, let's consider the perhaps most famous quantity implicature: the one involving the lexical contrast between "some" and "all". The utterance of (8), if compared with the alternative (9), may give rise either to the general epistemic implicature in (10a), the strong epistemic implicature in (10b), the weak epistemic implicature in (10c) or the factual implicature in (10d).

(8) Some of Kiki's friends are metalheads. "some"

(9) All of Kiki's friends are metalheads. "all"

(10) a. The speaker does not believe that all of Kiki's friends are metalheads. $\neg \text{Bel}_S$"all"

b. The speaker believes that not all of Kiki's friends are metalheads. $\text{Bel}_S \neg$"all"

c. The speaker is uncertain whether all of Kiki's friends are metalheads. Uc_S"all"

d. It's not the case that *all* of Kiki's friends are metalheads. $\neg$"all"

[7] This can be taken to higher-order reasoning as well, for the hearer could also exclude an alternative from consideration because he believes that, although the speaker is aware of it, the speaker believes that the hearer is not, etc.

Disjunction. An utterance of a plain disjunction as in (11) is usually associated with the *ignorance implicature* in (12) that the speaker is uncertain about each disjunct.

(11) Martha is in love with Alf or Bert. $A \vee B$

(12) a. The speaker is uncertain whether Martha is in love with Alf. $\mathrm{Uc}_S A$

b. The speaker is uncertain whether Martha is in love with Bert. $\mathrm{Uc}_S B$

There is no variation in the strength of the ignorance implicature, i.e., there are no further distinctions possible between general, strong or weak epistemic implicatures. This is so, because if the speaker believes that $A \vee B$ is true, but if she is also not uncertain about, say, A, then she must either know A or B.

Additionally, an utterance of (11) may give rise to an *exclusivity implicature* as in (13), which may arise as a factual or either of the three varieties of epistemic implicatures.

(13) a. The speaker does not believe that Martha is in love with both Alf and Bert. $\neg \mathrm{Bel}_S(A \wedge B)$

b. The speaker believes that Martha is not in love with both Alf and Bert. $\mathrm{Bel}_S \neg (A \wedge B)$

c. The speaker is uncertain whether Martha is in love with both Alf and Bert. $\mathrm{Uc}_S(A \wedge B)$

d. Martha is not in love with both Alf and Bert. $\neg(A \wedge B)$

It is often assumed that an utterance of (11) should be compared to the alternatives in (14) (cf. Sauerland 2004; Fox 2007).

(14) a. Martha is in love with Alf. A

b. Martha is in love with Bert. B

c. Martha is in love with Alf and Bert. $A \wedge B$

However, I am uncertain whether the conjunction in (14c) should always be included. It is also not self-evident to me that the exclusivity implicature, for which the conjunctive alternative in (14c) is needed, is always warranted. Often the exclusivity implicature seems to be due to world knowledge: when disjuncts are logically inconsistent, or when conjoined truth is highly unlikely (cf. Geurts 2010, for a similar sentiment).

Free Choice Disjunction. Intuitively, an utterance of a sentence like (15a) where a disjunction scopes under an existential modal operator has a so-called *free-choice implicature* as in (15b).[8]

(15) a. You may take an apple or a pear. $\Diamond(A \vee B)$

b. You may take an apple and you may take a pear. $\Diamond A \wedge \Diamond B$

The inference from (15a) to (15b) is not valid under a standard logical semantics which treats disjunction as the usual Boolean connective and the modal as an existential quantifier over accessible worlds. In Gricean spirit, we would like to account for this inference as a conversational implicature, indeed a factual quantity implicature.[9] One of the arguments in favor of an implicature-based analysis is the observation that the inference in (15b) seems to rest on the contextual assumption that the speaker is, in a sense, an authority about the deontic modality in question. If this assumption is not warranted or explicitly suspended as in example (16) we get the *ignorance implicatures* in (17).[10]

(16) You may take an apple or a pear, but I don't know which.

(17) a. The speaker is uncertain whether the hearer may take an apple. $\mathrm{Uc}_S(\Diamond A)$

b. The speaker is uncertain whether the hearer may take a pear. $\mathrm{Uc}_S(\Diamond B)$

Moreover, an utterance of (15a) may also receive *exclusivity implicatures* of various strength, as in (18).

(18) a. The speaker does not believe that the hearer may take both. $\neg\mathrm{Bel}_S\Diamond(A \wedge B)$

b. The speaker believes that the hearer may not take both. $\mathrm{Bel}_S\neg\Diamond(A \wedge B)$

c. The speaker is uncertain whether the hearer may not take both. $\mathrm{Uc}_S\Diamond(A \wedge B)$

[8]The problem has first explicitly been adressed in this form by Hans Kamp (Kamp 1973, 1978), but it is structurally related to "Ross's paradox" (cf. Ross 1941; Wright 1968). More recently, the problem of free-choice permission has inspired many interesting attempts of a solution. Semantically oriented approaches reconsider the semantics of disjunctions (Zimmermann 2000; Geurts 2005; Simons 2005) or of the modals involved (Merin 1992; van Rooij 2000; Asher and Bonevac 2005; Barker 2010). Pragmatic approaches, more or less close in spirit to Grice's programme, have been given by others (Kratzer and Shimoyama 2002; Chierchia 2004; Schulz 2005; Fox 2007).

[9]See also Kratzer and Shimoyama (2002); Alonso-Ovalle (2005); Schulz (2005) for more arguments why the free-choice inference in (15a) should be treated as an implicature.

[10]Other epistemic variants of ignorance implicatures do not arise. The case is entirely parallel to plain disjunctions. If the speaker believes that $\Diamond(A \vee B)$ is true and if she is not uncertain about $\Diamond A$ ($\Diamond B$), then she knows $\Diamond B$ ($\Diamond A$).

d. The hearer may not take both. $\neg\Diamond(A \wedge B)$

Again, as with plain disjunctions, I remain doubtful whether these latter exclusivity implicatures always arise and should systematically be derived as a quantity implicature by comparison with a conjunctive alternative. If I tell my students:

(19) You may consult your notes or your textbook during the exam.

it seems to me that it requires emphatic stress on "or" to convey that consulting both notes and textbook is not an option.

This ties in with the question which alternatives are operative in the derivation of these implicatures. As before with plain disjunctions, I will assume that an utterance of (15a) invites comparison with at least (20a) and (20b), and possibly, but not necessarily also (20c).

(20) a. You may take an apple. $\Diamond A$

b. You may take a pear. $\Diamond B$

c. You may take an apple and a pear. $\Diamond(A \wedge B)$

Simplification of Disjunctive Antecedents. The last case that I would like to consider concerns the interpretation of disjunction in the antecedents of conditionals. I will argue that we should preferably treat these in parallel with free-choice readings of disjunctions under existential modals (cf. van Rooij 2010).

Intuitively, the indicative conditional in (21) seems to convey both (22a) and (22b), and analogously the counterfactual conditional (23) seems to convey both (24a) and (24b).[11]

(21) If you eat an apple or a pear, you will feel better. $(A \vee B) \boxRightarrow C$

(22) a. If you eat an apple, you will feel better. $A \boxRightarrow C$

b. If you eat a pear, you will feel better. $B \boxRightarrow C$

(23) If you'd eaten an apple or a pear, you'd feel better. $(A \vee B) \boxRightarrow C$

(24) a. If you'd eaten an apple, you'd feel better. $A \boxRightarrow C$

b. If you'd eaten a pear, you'd feel better. $B \boxRightarrow C$

[11] I make use of the following notation: $A \boxRightarrow B$ is an abbreviation for "if A, then will/would B" and $A \Diamond\!\!\Rightarrow B$ stands for "if A, then may/might B", irrespective of whether the conditional is in the subjunctive of the indicative mood. I write $A > B$ for any conditional with either universal or existential modal in the consequent.

Similar inferences are warranted for conditionals with existential modals in their antecedents: an utterance of (25) may convey both (24a) and (24b); an utterance of (27) may convey both (26a) and (26b).

(25) If you eat an apple or a pear, you might feel better. $(A \vee B) \diamond\!\!\Rightarrow C$

(26) a. If you eat an apple, you might feel better. $A \diamond\!\!\Rightarrow C$

b. If you eat a pear, you might feel better. $B \diamond\!\!\Rightarrow C$

(27) If you'd eaten an apple or a pear, you might feel better. $(A \vee B) \diamond\!\!\Rightarrow C$

(28) a. If you'd eaten an apple, you might feel better. $A \diamond\!\!\Rightarrow C$

b. If you'd eaten a pear, you'd feel better. $B \diamond\!\!\Rightarrow C$

In general, the inference from $(A \vee B) > C$ to $A > C$ (or $B > C$) is known as SIMPLIFICATION OF DISJUNCTIVE ANTECEDENTS, henceforth SDA.

Although SDA is a valid inference under a material implication analysis of conditionals, standard possible-worlds semantics in the vein of Stalnaker (1968) and Lewis (1973) do not make SDA valid. According to these theories, we evaluate a conditional as true or false in a given possible w with respect to a set R_w of worlds accessible from w and some —let's assume: well-founded— ordering $\preceq_w$ on this set. Many reasonable constraints on the nature of this ordering could be given to instantiate certain influential theories of conditionals (think of: Stalnaker 1968; Lewis 1973; Kratzer 1981; Lewis 1981; Veltman 1985). For the present pragmatic purpose we should remain noncommittal and not take on *any* particular constraints on the ordering, except that it be well-founded, i.e., that it conforms to the *limit assumption*, so as to facilitate notation. We then first define

$$\mathrm{Min}(R_w, \preceq_w, A) = \{v \in R_w \cap A \mid \neg\exists v' \in R_w \cap A \ : \ v' \prec_w v\}$$

and then define for indicative or counterfactual conditionals alike that:

$$A \boxdot\!\!\Rightarrow C \text{ is true in } w \quad \text{iff} \quad \mathrm{Min}(R_w, \preceq_w, A) \subseteq C\,,$$

as well as:

$$A \diamond\!\!\Rightarrow C \text{ is true in } w \quad \text{iff} \quad \mathrm{Min}(R_w, \preceq_w, A) \cap C \neq \emptyset\,.$$

With this, it is plain to see that SDA is not a generally valid inference under these semantics, because, for instance, $A \vee B \boxdot\!\!\Rightarrow C$ could be true if all minimal worlds where $A \vee B$ is true are such that A and C are true and B is false, while all minimal B-worlds are worlds where C is false. In that case $(A \vee B) \boxdot\!\!\Rightarrow C$ would be true but $B \boxdot\!\!\Rightarrow C$ would be false. Whence that SDA is not semantically valid.

Should we worry? Some say yes, some say no. From those who say yes, the invalidity of SDA has been held as a problem case against in particular Lewis's (1973) theory of counterfactuals (see Nute 1975; Fine 1975), but the case would equally apply to indicatives under like-minded semantic theories. On the other hand, there are good arguments not to want SDA to be a semantically valid inference pattern. Warmbrōd (1981) gives a strong argument in favor of this position. He argues that if a conditional semantics makes SDA valid, and if we otherwise stick to standard truth-functional interpretation of disjunction, we can also derive that inferences like that from (29a) to (29b) are generally valid, which intuitively should not be the case.[12]

(29) a. If you eat an apple, you will feel better. $A \boxdot\!\!\Rightarrow C$

b. If you eat an apple and a rock, you'll feel better. $(A \wedge B) \boxdot\!\!\Rightarrow C$

Another argument against a semantic validation of SDA comes from examples such as the following (cf. McKay and van Inwagen 1977):

(30) a. If John had taken an apple or a pear, he would have taken an apple.

b. If John had taken a pear, he would have taken an apple.

If SDA was semantically valid then (30a) would imply (30b), but this is of course nonsense. Together, this suggests loosely that SDA should perhaps be thought of as a pragmatic inference on top of a standard semantics, which is what I argue for here.[13]

Indeed, it has been noted before that SDA looks strikingly similar to free choice permission in several respects (cf. Klinedinst 2006; van Rooij 2006). Firstly, English conditionals like (31) *can* be used to grant permissions.

(31) It's fine with me if you take an apple.

This may not be the most frequent and natural way of giving permissions, but it seems that at least the question

(32) Is it okay if I take an apple?

is an expression frequently used to *ask* for permission in English. This is different in other languages though. For instance, lacking a clear equivalent to English modal "may", a standard construction for permission giving in Japanese is the conditional construction "-te mo" which generally translates as "even if" (see McClure 2000, p. 180):

[12]Formally, this is because if SDA is generally valid, we can infer from $A \boxdot\!\!\Rightarrow C$ and the fact that $(A \wedge B) \vee (A \wedge \neg B)$ is a truth-functionally equivalent to A that $((A \wedge B) \vee (A \wedge \neg B)) \boxdot\!\!\Rightarrow C$. Then, by SDA, we derive $(A \wedge B) \boxdot\!\!\Rightarrow C$ for arbitrary B.

[13]So here I depart slightly from ultra-orthodox Griceanism, if you want to call it that: I consider (something like) the above order-sensitive possible worlds semantics the standard, and I do not consider material implication a respectable semantic analysis of conditionals.

(33) ringo wo tabe-te mo ii.
apple Object Marker eat-TE-Form also good
'It's good even if you eat an apple.'
'You may eat an apple.'

Most crucially, a conditional similar to (31) with a disjunctive antecedent as in (34) is certainly taken to convey that both sentences in (35) are true.

(34) It's fine with me if you take an apple or a pear.

(35) a. It's fine with me if you take an apple.

b. It's fine with me if you take a pear.

Moreover, we can force epistemic ignorance readings also for conditionals with disjunctive antecedents (cf. Klinedinst 2006), quite parallel to cases like (16)–(17):

(36) a. If you eat an apple or a pear, you will feel better, but I don't know which. $(A \vee B) \boxdot\!\!\Rightarrow C$

b. The speaker is uncertain whether $A \boxdot\!\!\Rightarrow C$ is true. $\mathrm{Uc}_S(A \boxdot\!\!\Rightarrow C)$

c. The speaker is uncertain whether $B \boxdot\!\!\Rightarrow C$ is true. $\mathrm{Uc}_S(B \boxdot\!\!\Rightarrow C)$

In a context like (36a) that marks the speaker's epistemic uncertainty we do not take an utterance of $(A \vee B) \boxdot\!\!\Rightarrow C$ that the speaker *knows* that $A \boxdot\!\!\Rightarrow C$ and $B \boxdot\!\!\Rightarrow C$ are both true, as we should if SDA was a semantically valid inference. Rather, if we take the speaker to be maximally knowledgeable despite her expressed uncertainty, we only infer that the speaker considers exactly one of the sentence $A \boxdot\!\!\Rightarrow C$ and $B \boxdot\!\!\Rightarrow C$ true, without knowing which one this is.

Taken together, SDA looks remarkably similar to free choice permission and we should try and see whether SDA can be derived as a pragmatic inference similar to free choice readings under existential modals. This is the general point of view which I will adopt in the following. More concretely, I would like to propose that it is desirable —in the classical Gricean sense of appealing to parsimony— to account for the behavior of disjunction under existential modals on a par with SDA. Interestingly, our currently best formal models of quantity implicatures fail to do so, as the following section will show.

4 Exhaustive Interpretation

The currently most prominent formal approaches to quantity implicatures are formulated in terms *exhaustivity operators*. These were originally conceived to account for the exhaustive reading of answers to questions (cf. Groenendijk

and Stokhof 1984; von Stechow and Zimmermann 1984). When it comes to quantity implicatures, there are two conceptually and formally distinct versions of exhaustivity operators, one is more semantic, the other is more syntactic in nature. The semantic approach —worked out, *inter alia*, by van Rooij and Schulz (2004), Schulz and van Rooij (2006) and Spector (2006)— tries to account for quantity implicature as interpretation in *minimal models*. That is why I refer to it as a semantic approach here. This is in contrast with an approach to exhaustive interpretation in terms of a notion called *innocent exclusion* —due to Fox (2007) and taken over in some form or other by, among others, Alonso-Ovalle (2008) and Chierchia et al. (2008). The latter is not formulated in terms of possible worlds, but rather in terms of explicit reasoning about alternatives. That is why I refer to it as the syntactic approach.

The following briefly introduces both of these approaches in sufficient formal detail for further comparison. As for notation, let S be the sentence uttered in whose interpretation we are interested in and let ALT be the set of contextually associated alternatives (with $S \in$ ALT).[14]

4.1 The minimal-models approach.

The general idea of the minimal-models approach to exhaustive interpretation is to define an ordering on possible worlds (or the speaker's information states) based on the available alternatives ALT and to consider as pragmatic interpretation of a given sentence all the worlds (or states) that are minimal with respect to this ordering. The relevant ordering is obtained from ALT by defining a world (or state) to be more minimal than another if strictly more alternatives from ALT are false (not believed true). Its proponents suggest that in this way the minimal-models approach to exhaustive interpretation captures pragmatic interpretation based on (something like) the Speaker Quantity Principle (3), because the interpretation selects those worlds or states where fewest possible alternatives are true, because, in turn, a cooperative speaker may be expected to have uttered these alternatives, if she had had the relevant knowledge. (We will see below that this characterization of the formal operation is off the mark: exhaustive interpretation is best conceived as rational hearer interpretation without necessarily taking a Gricean cooperative speaker into account.)

[14]Here and in the following, I will be sloppy in using capital letters as variables for both sentences and the propositions (sets of possible worlds) that they denote. To make matters *even worse*, I happily administer logical notation to the mix: for instance, $\neg S$ would be the negation of sentence S, or the complement of set S in the set of all possible worlds.

Factual Implicatures. To account for factual implicatures, we consult a partial ordering on possible worlds $<_{\textsc{Alt}}$, defined as follows:

$$w' <_{\textsc{Alt}} w \qquad \text{iff} \qquad \{A \in \textsc{Alt} \mid w' \in A\} \subset \{A \in \textsc{Alt} \mid w \in A\} \; .$$

The standard exhaustive interpretation of S —which captures factual quantity implicatures— is obtained from the assumption that the actual world is minimal with respect to this order in the set of worlds where S is true:

$$\textsc{Exh}_{\text{MM}}(S, \textsc{Alt}) = \{w \in S \;\mid\; \neg\exists w' \in S \;:\; w' <_{\textsc{Alt}} w\} \; . \tag{4.1}$$

For example, take a simple scalar contrast between sentences, "some A's are B's" and "all A's are B's" (such as between (8) and (9)). Then we need to consider only two types of possible worlds: $w_{\exists\neg\forall}$ where some but not all A's are B's, and $w_{\forall}$ where all A's are B's. The ordering defined above gives us: $w_{\exists\neg\forall} <_{\textsc{Alt}} w_{\forall}$, because in $w_{\forall}$ both alternative sentences are true, while in $w_{\exists\neg\forall}$ only the weaker "some" statement is. Consequently, the pragmatic interpretation according to this approach selects:

$$\textsc{Exh}_{\text{MM}}(\text{"some } A\text{'s are } B\text{'s"}, \textsc{Alt}) = \{w_{\exists\neg\forall}\} \; .$$

Notice that $w_{\exists\neg\forall}$ is a stand-in for a class of possible worlds: the pragmatic interpretation of the sentence is that some but not all A's are B's, just as we want it to.

Epistemic Implicatures. To account for epistemic implicatures in terms of minimal models, we consult an ordering $<^{\Box}_{\textsc{Alt}}$ defined, not on possible worlds, but on information states of the speaker.[15] As usual, an information state s is a non-empty set of possible worlds, collecting the possibilities that cannot be ruled out for certain by, in our case, the speaker. To compute general epistemic implicatures, information states are compared with respect to how many proposition from Alt the speaker knows to be true:

$$s' <^{\Box}_{\textsc{Alt}} s \qquad \text{iff} \qquad \{A \in \textsc{Alt} \mid s' \subseteq A\} \subset \{A \in \textsc{Alt} \mid s \subseteq A\} \; .$$

Based on this ordering, a cooperative speaker who utters S is predicted to be in a state in the set $\textsc{Exh}^{\text{GE}}_{\text{MM}}(S, \textsc{Alt})$ defined as follows:

$$\textsc{Exh}^{\text{GE}}_{\text{MM}}(S, \textsc{Alt}) = \{s \subseteq S \mid \neg\exists s' \subseteq S \;:\; s' <^{\Box}_{\textsc{Alt}} s\} \; . \tag{4.2}$$

[15] We could still use an ordering on possible worlds if these are associated with an information state of the speaker, but this is not needed for anything of relevance here and would only complicate notation. My exposition here follows the variant of exhaustive interpretation of van Rooij and Schulz (2006).

These are then all the information states where fewest alternatives are *believed true* within the class of information states where the target sentence S believed true.

To see how this accounts for the general epistemic quantity implicature of cases like (8), i.e., the comparison between "some A's are B's" and "all A's are B's", we need to consider all information states expressible in terms of the types of worlds $w_{\exists\neg\forall}$ and $w_{\forall}$ from above. These are: $s_1 = \{w_{\exists\neg\forall}, w_{\forall}\}$ where the speaker is uncertain, $s_2 = \{w_{\exists\neg\forall}\}$ where the speaker knows that some but not all A's are B's, and $s_3 = \{w_{\forall}\}$ where the speaker knows that all A's are B's. The ordering $<^{\Box}_{\textsc{Alt}}$ as defined above yields: $s_1, s_2 <^{\Box}_{\textsc{Alt}} s_3$. With this, we derive:

$$\textsc{Exh}^{\textsc{GE}}_{\textsc{MM}}(\text{"some } A\text{'s are } B\text{'s"}, \textsc{Alt}) = \{s_1, s_2\} \ ,$$

which says that the speaker is in an information state where she does not believe that all A's are B's.

This prediction can be strengthened to obtain the strong epistemic implicatures if we assume that the speaker is competent. Following van Rooij and Schulz, this can be expressed by layering another ordering $<^{\Diamond}_{\textsc{Alt}}$ on the outcome of interpretation $\textsc{Exh}^{\textsc{GE}}_{\textsc{MM}}(S, \textsc{Alt})$. This ordering $<^{\Diamond}_{\textsc{Alt}}$ ranks information states with respect to which alternatives from A are considered possible:

$$s <^{\Diamond}_{\textsc{Alt}} s' \quad \text{iff} \quad \{A \in \textsc{Alt} \mid s \cap A \neq \emptyset\} \subset \{A \in \textsc{Alt} \mid s' \cap A \neq \emptyset\} \ .$$

Using this ordering, we can define an exhaustivity operator to capture strong epistemic implicatures as follows:

$$\textsc{Exh}^{\textsc{SE}}_{\textsc{MM}}(S, \textsc{Alt}) = \{s \in \textsc{Exh}^{\textsc{GE}}_{\textsc{MM}}(S, \textsc{Alt}) \mid \neg\exists s' \in \textsc{Exh}^{\textsc{GE}}_{\textsc{MM}}(S, \textsc{Alt}) \ : \ s' <^{\Diamond}_{\textsc{Alt}} s\}$$

In the example above, we get $s_2 <^{\Diamond}_{\textsc{Alt}} s_1$, so that:

$$\textsc{Exh}^{\textsc{SE}}_{\textsc{MM}}(\text{"some } A\text{'s are } B\text{'s"}, \textsc{Alt}) = \{s_2\} \ .$$

The strong epistemic implicature that the speaker believes that some but not all A's are B's is predicted. (The minimal-models approach does not attend in detail to the difference between general and weak epistemic implicatures.)

4.2 The innocent-exclusion approach.

Fox (2007) pursues a different approach to quantity implicatures via exhaustive interpretation. According to Fox, factual quantity implicatures are computed as part of the syntactic system by an exhaustivity operator that is defined differently from the above. Epistemic implicatures are computed, according to Fox, in the more traditional Gricean manner in the vein of Sauerland (2004). This division of labor between syntactic and pragmatic systems

calculating different types, or, if you wish, aspects of quantity implicatures is unappealing from a methodological point of view (cf. Geurts 2010). But be that as it may, for our present purposes it is important to notice that Fox's alternative approach to the base-level exhaustive interpretation operator appears better suited to deal with the data than the minimal-models approach.

Fox's definition of the exhaustivity operator makes use of a novel notion which he calls *innocent exclusion*. The idea is that scalar implicatures of S are computed by negating all alternatives that can be excluded *consistently* without making an *arbitrary* choice in excluding these.

Towards a definition of innocent exclusion, first define what it means to be *consistently excludable*. A subset A of $\textsc{Alt}$ is consistently excludable if negating all elements in A is consistent with the truth of S:

$$\text{CE}(S, \textsc{Alt}) = \left\{ X \subseteq \textsc{Alt} \mid \bigwedge_{A \in X} \neg A \text{ is consistent with } S \right\} .$$

We would like to exclude as many of the alternatives as possible. So we should look at the set of maximal elements in $\text{CE}(S, \textsc{Alt})$:

$$\text{Max-CE}(S, \textsc{Alt}) = \{X \in \text{CE}(S, \textsc{Alt}) \mid \neg \exists Y \in \text{CE}(S, \textsc{Alt}) \ : \ X \subset Y\} .$$

So, for each set of alternatives $X \in \text{Max-CE}(S, \textsc{Alt})$, negating all elements $A \in X$ would be a maximally consistent pragmatic enrichment of the target sentence S. But it may be the case that some alternatives $A \in \textsc{Alt}$ occur only in some but not all of the maximal sets in $\text{Max-CE}(S, \textsc{Alt})$. Excluding these would be an *arbitrary* choice. That's why Fox defines the set of innocently excludable alternatives $\text{IE}(S, \textsc{Alt})$ to S given $\textsc{Alt}$ as those alternatives that are in *every* maximally consistent pragmatic enrichment:

$$\text{IE}(S, \textsc{Alt}) = \bigcap \text{Max-CE}(S, \textsc{Alt}) .$$

Exhaustive interpretation based on innocent exclusion is then defined straightforwardly:

$$\textsc{Exh}_{\text{IE}}(S, \textsc{Alt}) = S \wedge \bigwedge_{A \in \text{IE}(S, \textsc{Alt})} \neg A .$$

How does this definition differ from the minimal-models approach? Generally the operators $\textsc{Exh}_{\text{MM}}$ and $\textsc{Exh}_{\text{IE}}$ are not equivalent. The precise formal relation is worked out in Appendix A. In a nutshell, it turns out that exhaustification based on innocent exclusion is always subsumed under the minimal models interpretation, but that the latter may rule out worlds that the former doesn't, depending on the set of explicit alternatives. This property also matters for explanations of the basic free choice implicature which, for illustration, I will briefly go through next.

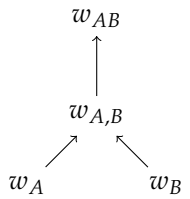


Figure 1: Ordering on worlds based for basic free choice implicature

Predictions for Basic Free Choice Disjunctions. Recall that we are interested in explaining how (15a) can have the factual implicature (15b), repeated here.

(15a) You may take an apple or a pear. $\Diamond(A \vee B)$

(15b) You may take an apple and you may take a pear. $\Diamond A \wedge \Diamond B$

If we take the set of alternatives $\textsc{Alt} = \{\Diamond A,\ \Diamond B,\ \Diamond(A \wedge B),\ \Diamond(A \vee B)\}$, then according to the minimal worlds approach we have to distinguish four kinds of worlds: (i) in w_A $\Diamond A$ is true but $\Diamond B$ is not, (ii) in w_B $\Diamond B$ is true but $\Diamond A$ is not, (iii) in $w_{A,B}$ both $\Diamond A$ and $\Diamond B$ are true, but $\Diamond(A \wedge B)$ is false, and, finally, (iv) in w_{AB} $\Diamond(A \wedge B)$ is true. The ordering $<_{\textsc{Alt}}$ is a partial order as depicted in Figure 1. (An arrow from w to w' symbolizes $w <_{\textsc{Alt}} w'$; arrows that follow from transitivity are left out for readability.) The minimal worlds according to that ordering are:

$$\textsc{Exh}_{\text{MM}}(\Diamond(A \vee B), \textsc{Alt}) = \{w_A, w_B\}\ .$$

This prediction is too strong, for it actually rules out the free-choice implicature in (15b). The free-choice implicature would correspond to an interpretation that selects $\{w_{A,B}\}$ as the pragmatic interpretation for our target sentence.

Innocent exclusion is weaker on this set of alternatives and, crucially, does *not* exclude the free-choice reading. In this case, there are two maximal sets of consistently excludable alternatives:

$$\text{Max-CE}(\Diamond(A \vee B), \textsc{Alt}) = \{\{\Diamond A,\ \Diamond(A \wedge B)\},\ \ \{\Diamond B,\ \Diamond(A \wedge B)\}\}\ .$$

The intersection of these contains only $\Diamond(A \wedge B)$, so that:

$$\textsc{Exh}_{\text{IE}}(\Diamond(A \vee B), \textsc{Alt}) = \{w_A, w_B, w_{A,B}\}\ .$$

In sum, $\textsc{Exh}_{\text{IE}}$ does not rule out the free-choice inference, but it does not predict it either.

Iterated Exhaustification. This is why Fox (2007) suggests to apply the exhaustification operator, if needed, several times, so as also to factor in the exhaustive interpretation of the alternatives. The repeated application of exhaustification based on innocent exclusion is defined as follows:[16]

$$\begin{aligned}
\textsc{Alt}_1 &= \textsc{Alt} \\
\textsc{Exh}_1(A) &= \textsc{Exh}_{\textsc{IE}}(A, \textsc{Alt}_1) \\
\textsc{Alt}_{n+1} &= \{\textsc{Exh}_n(A) \mid A \in \textsc{Alt}_n\} \\
\textsc{Exh}_{n+1}(A) &= \textsc{Exh}_{\textsc{IE}}(\textsc{Exh}_n(A), \textsc{Alt}_{n+1})
\end{aligned}$$

Fox shows that further iteration of the exhaustification operator derives the basic free choice reading. I briefly repeat Fox's result here. Establish first the non-iterated exhaustive readings of all alternatives:

$$\begin{aligned}
\textsc{Exh}_1(\Diamond A) &= \Diamond A \wedge \neg\Diamond B \\
\textsc{Exh}_1(\Diamond B) &= \Diamond B \wedge \neg\Diamond A \\
\textsc{Exh}_1(\Diamond(A \wedge B)) &= \Diamond(A \wedge B) \\
\textsc{Exh}_1(\Diamond(A \vee B)) &= \Diamond(A \vee B) \wedge \neg\Diamond(A \wedge B)
\end{aligned}$$

This is then the new set of alternatives $\textsc{Alt}_2$ for input into another round of exhaustification. This will affect only the target sentence, whose new interpretation becomes:

$$\begin{aligned}
\textsc{Exh}_2(\Diamond(A \wedge B)) &= \textsc{Exh}_1(\Diamond(A \vee B)) \wedge \neg\textsc{Exh}_1(\Diamond A) \wedge \neg\textsc{Exh}_1(\Diamond B) \\
&= \{w_{A,B}\} \ .
\end{aligned}$$

Voila, free choice derived!

It is worthwhile to mention that in the case under consideration here already after one additional round of iterated exhaustification the interpretation has reached a *fixed point*: all subsequent iterations will yield the same outcome. It is easy to check, and in fact Spector (2007) provides a proof for his version of iterated exhaustification based on minimal models, that for finite set $\textsc{Alt}$ the system must reach a fixed point after finitely many iteration steps. This follows from the simple observation that iterated exhaustification is a monotone operation: $\textsc{Exh}_{n+1}(A) \subseteq \textsc{Exh}_n(A)$ for all n.

[16] Spector (2007) utilises a parallel definition of iterated exhaustification in terms of the minimal models approach. It is easy to verify that this will not help to derive the basic free-choice implicature, as the desired reading is already excluded at the first step and moreover, $\textsc{Exh}_{n+1}(A) \subseteq \textsc{Exh}_n(A)$ for all A and n.

Problems with Iterated Exhaustification. Fox's proposal neatly accounts for the basic free-choice reading of sentences like (15a), but it does have its problems too. Firstly, as already mentioned, some would certainly prefer a uniform account for factual and epistemic implicatures. Secondly, although the minimal-models approach can be traced, albeit somewhat vaguely perhaps, to a Gricean rationale, it is not clear where exactly the notion of *innocent exclusion* comes from. From a Gricean perspective, we would certainly like to see a further grounding of this notion in terms of rationality. Thirdly, there are also empirical problems with the innocent-exclusion approach. Here is one of them.

Despite the superficial parallel between free-choice implicatures and SDA, Fox's system does not predict SDA, if we assume a standard Lewis-Stalnaker semantics for conditionals as outlined in Section 4. The problem in a nutshell is that although $\Diamond(A \vee B)$ is entailed by, say, $\Diamond A$, and with this the basic free-choice reading can be derived, it is not the case that $(A \vee B) > C$ is entailed by $A > C$. This precludes SDA to be derived by iterated innocent exclusion.

Making things more concrete, reconsider the target sentence (21), which we take to implicate that both (22a) and (22b) are true.

(21)	If you eat an apple or a pear, you will feel better.	$(A \vee B) \boxRightarrow C$
(22a)	If you eat an apple, you will feel better.	$A \boxRightarrow C$
(22b)	If you eat a pear, you will feel better.	$B \boxRightarrow C$

For simplicity, it suffices to consider the set of alternatives (the argument also holds if we include the conjunctive alternative):

$$\textsc{Alt} = \{A \boxRightarrow C,\ B \boxRightarrow C,\ (A \vee B) \boxRightarrow C\}\ .$$

In a first round of applying the exhaustivity operator $\textsc{Exh}_{\text{IE}}$ to our target sentence we cannot exclude any alternatives innocently, because if $(A \vee B) \boxRightarrow C$ is true, then at least one of $A \boxRightarrow C$ or $B \boxRightarrow C$ must be true as well:

$$\textsc{Exh}_1((A \vee B) \boxRightarrow C) = (A \vee B) \boxRightarrow C\,.$$

This is as with the basic free-choice inference we calculated before. But we now get different prediction for the other alternatives, because, the maximal sets of consistently excludable alternatives to $A \boxRightarrow C$ and $B \boxRightarrow C$ both contain the target sentence:

$$\textsc{Exh}_1(A \boxRightarrow C) = A \boxRightarrow C \wedge \neg B \boxRightarrow C \wedge \neg((A \vee B) \boxRightarrow C)$$
$$\textsc{Exh}_1(B \boxRightarrow C) = B \boxRightarrow C \wedge \neg A \boxRightarrow C \wedge \neg((A \vee B) \boxRightarrow C)\,.$$

But that means that at the next application of exhaustification of $(A \vee B) \boxRightarrow C$ neither $\textsc{Exh}_1(A \boxRightarrow C)$ nor $\textsc{Exh}_1(B \boxRightarrow C)$ can be consistently negated! We

have already reached a fixed point without enriching the target sentence to support SDA.

In conclusion, the exhaustive-interpretation approach to quantity implicatures seems promising, but is not without its own problems either. Most importantly, we do not get a uniform explanation for choice-readings of disjunctions and SDA. This calls for a general account, rooted in the assumption that interlocutors behave rationally towards the common shared goal of conversation, that (i) accounts for the free choice-readings of disjunctions, SDA, as well as the all the epistemic implicatures, in a uniform way, and that (ii) ideally also sheds light on the precise nature of (iterated) exhaustive interpretation. This is what the following game theoretic model does. In order to motivate its general set-up —an explanation of quantity implicatures in terms of rational behavior—, I would first like to briefly discuss what is normally considered a rational explanation of behavior in general and of conversational implicatures in particular.

5 Rationalizing Quantity Implicatures

If Grice's conjecture about a possible rational foundation of implicatures is correct, then a quantity implicature is really an *abductive inference* that rationalizes why the speaker has made a certain utterance (cf. Geurts 2010). In simple cases, this inference is still rather perspicuous. It is a piece of hearer reasoning that departs, or so I propose, from the following premises:

(P 1) The speaker uttered φ.

(P 2) The speaker could have uttered ψ also and was aware of that.

(P 3) The speaker rationally chose φ over ψ for a reason.

What we would like to conclude from this is at least the general epistemic implicature:

(C) The speaker does not believe that ψ is true.

which could, of course, be strengthened by competence reasoning where necessary. In what sense and under which circumstances does this conclusion follow from these premisses?

The pivotal element clearly is premisse 3, the speaker's rationality. A choice of φ over ψ is rational if the speaker believes that φ suits her purpose no worse than a choice of ψ. But that means that a choice can be rational for a sheer myriad of reasons. With some vivid fantasy we may concoct ever

new pairs of beliefs and preferences to ascribe to the speaker, all of which could make her choice a rational one.

This is where the abductive nature of the inference surfaces most clearly: this is not a deduction, but an inference to the best explanation. What a "best explanation" is in a given situation, is not a matter of logical necessity but of common sense. What we are looking for, then, is a set of plausible extra assumptions X about the speaker's mental state, her beliefs and preferences, that satisfies two conditions: (i) the general epistemic implicature should be contained in X, and (ii) the conjunction of X should (defeasibly and non-trivially) entail the speaker's rationality. For instance, the general epistemic implicature in (C) entails the speaker's rationality in conjunction with the following two assumptions about speaker beliefs (A 1) and preferences (A 2):

(A 1) The speaker believes that the hearer will come to believe that ψ is true if she utters it.

(A 2) The speaker wants the hearer to believe a proposition only if she believes it too.

The problem for a systematic grounding of quantity implicatures in terms of rationality is now clear: we would need more clarity concerning additional assumptions about the speaker's mental state; ideally, we would like to have a perspicuous model of interactive speaker and hearer beliefs and preferences that is derived from a handful of innocuous and plausible assumptions about how interlocutors may construct a representation of the context of utterance when needed, and also of the beliefs that interlocutors have about each others' behavior and beliefs. This is why we should turn to *game theory*, which gives us exactly what we need: (i) a sufficiently detailed and principled context-model, (ii) a formal notion of rationality in terms of agents' beliefs and preferences, and (iii) a systematic way of assessing players' interactive beliefs about beliefs and action choices.[17]

6 Interpretation Games

An utterance and its uptake can be represented in terms of a *signaling game*. Signaling games have been studied extensively in philosophy (Lewis 1969), linguistics (e.g. Parikh 2001; van Rooij 2004; Jäger 2007), biology (e.g. Grafen 1990) and economics (e.g. Spence 1973). A signaling game is a simple dynamic game with imperfect information between two players: a sender and a receiver. The sender knows the actual state of the world t, but the receiver

[17]For general introduction to game theory in the context of linguistic pragmatics, see Benz et al. (2006).

doesn't. The sender chooses a message m from a given set of alternatives, all of which we assume here to have a semantic meaning commonly known between players. The receiver observes the sent message m and chooses an action a. An *outcome* of playing a signaling game for one round is given by the triple t, m and a. Each player has his own preferences over such outcomes.

Formally, a signaling game (with meaningful signals) is a tuple

$$\langle \{S, R\}, T, \Pr, M, [\![\cdot]\!], A, \mathrm{U}_S, \mathrm{U}_R \rangle$$

where sender S and receiver R are the players of the game; T is a set of states of the world; $\Pr \in \Delta(T)$ is a prior probability distribution over T, which represents the receiver's uncertainty which state in T is actual;[18] M is a set of messages that the sender can send; $[\![\cdot]\!] : M \to \mathcal{P}(T) \setminus \emptyset$ is a denotation function that gives the predefined semantic meaning of a message as the set of all states where that message is true; A is the set of response actions available to the receiver; and $\mathrm{U}_{S,R} : T \times M \times A \to \mathbb{R}$ are utility functions for both sender and receiver, mapping each outcome $\langle t, m, a \rangle$ to a numerical payoff that represents how desirable this outcome is to the player.

Towards an explanation of quantity implicatures in natural language, a special class of signaling games is of particular relevance: those which implement the basic Gricean assumptions of cooperativity and relevance of information. I will refer to signaling games that implement these assumptions as *interpretation games*. Interpretation games are representations of a context of utterance of a to-be-interpreted sentence that capture all and only the essential features of an utterance context that standardly trigger quantity reasoning. I suggest that these context representations are constructed generically from the usual set of alternatives to the to-be-interpreted expression, together with their logical semantics.

I will make a distinction between *base-level* and *epistemic* interpretation games. The former serve as representations of an utterance context of a hearer who is not consciously taking the speaker's epistemic states into account.[19] Base-level interpretation games are where we derive factual quantity implicatures. Epistemic interpretation games *do* explicitly accommodate the speaker's epistemic states and its these context models that we draw on to explain epistemic quantity implicatures.

[18] As for notation, $\Delta(X)$ is the set of all probability distributions over set X, Y^X is the set of all functions from X to Y, $X : Y \to Z$ is alternative notion for $X \in Z^Y$, and $\mathcal{P}(X)$ is the power set of X.

[19] I propose that this is natural and happens a lot: when a trusted source —think: your mother— says "I am proud of you" you often *directly* integrate the information "mother is proud of me" into your stock of beliefs without always reasoning about your mother's beliefs and authority on the issue at hand. (The latter we do only when forced by the context or otherwise necessary.)

6.1 Base-Level Interpretation Games

To construct a base-level interpretation game for the interpretation of a sentence S with alternatives ALT, we equate first the set ALT with the set of speaker available messages M. Next, we need to define a reasonable set of state distinctions T that are relevant for quantity reasoning. These distinctions hinge on the available alternatives. Clearly, not every possible way the world could be can be distinguished with any set M, and we should therefore restrict ourselves to only those states of affairs that can feasibly be expressed with the linguistic means at hand. Which are these exactly?

Given a set of propositions ALT and two possible worlds w and v, we say that w and v are ALT-indistinguishable, $w \sim_{\text{ALT}} v$, iff for all $A \in \text{ALT}$ we have $w \in A \Leftrightarrow v \in A$. Quantity reasoning for the interpretation of S based on ALT should look at all those worlds in which S is true, but we may safely lump together worlds that are ALT-indistinguishable. Consequently, the set of base-level state distinctions is:

$$T_{\text{BL}} = \{\{w \in S \mid w \sim_{\text{ALT}} v\} \mid v \in S\} \,.$$

Of course, the choice of semantic denotation function $[\![\cdot]\!]$ in our game model is then obvious. A message m_{A} is true in t, $t \in [\![m_{\text{A}}]\!]$, if $t \subseteq A$ for that alternative $A \in \text{ALT}$ that m_{A} represents in our context model.[20]

As for the prior probabilities $\Pr(\cdot)$, since we are dealing with general models of utterance interpretation, we would often not assume that the receiver has biased beliefs about which specific state obtains that are relevant for quantity reasoning.[21] In the absence of interpretation-relevant beliefs, we may make a simplifying assumption that $\Pr(\cdot)$ is a *flat probability distribution*:[22]

$$\Pr(t) = \Pr(t') \quad \text{for all } t, t' \in T \,.$$

Next, the set of receiver actions is equated with the set of states $A = T$ and the receiver's utilities model his interest in getting to know the true state of

[20] In the following, I will reuse this notation and write m_{A} for the message of the signaling game that corresponds one-to-one to alternative $A \in \text{ALT}$.

[21] See Allott (2006) and Franke (2009, Chapter 3) for discussion what prior probabilities in an interpretation game could and could not represent.

[22] This last assumption may seem contentious. In its defense I would like to say two things: (i) it is not strictly speaking necessary (cf. Franke 2009), but (ii) if we adopt this assumption we can simplify the system of pragmatic reasoning that I will introduce in Section 8.3 dramatically, so much so that we can stop entirely to care about probabilities, which, I believe, is a great relief to the working linguist who may not be too familiar with the details of information theory and probability reasoning. (But see also Geurts and Pouscoulous (2009*a*) for arguments that beliefs about likely worldly states of affairs often do not seem to impact quantity reasoning.)

affairs, i.e., getting the right *interpretation* of the observed message:

$$\mathrm{U}_R(t, m, a) = \begin{cases} 1 & \text{if } t = a \\ 0 & \text{otherwise.} \end{cases}$$

The assumption underlying this construction is that quantity reasoning is about coordinating which meaning enrichment of the target message is reasonable given a set of possible meaning distinctions induced by the alternatives. Let me briefly enlarge on this.

Any signaling game embeds the structure $\langle T, \Pr, A, \mathrm{U}_R \rangle$ which is a classical *decision problem* of the receiver. Following van Rooij (2003), we may look at such a decision problem as a generalization of the notion of a question. In the present case this gives a flexible and powerful representation of a *question under discussion*: the receiver's decision problem pins down which state distinctions matter in which way to the (practical) decision of the hearer. In other words, this part of the structure implements what is *relevant* to the hearer. Consequently, interpretation games assume that the hearer is interested in quantity reasoning, i.e., that he is interested in drawing any finer meaning distinctions that alternative messages could have made.

This settles the utilities of the receiver. As for the speaker, we should assume that conversation is a *cooperative* effort —at least on the level of such generic context models that back up quantity reasoning. This is easily implemented by defining the sender's utilities in terms of the receiver's as follows:[23]

$$\mathrm{U}_S(t, m, a) = \mathrm{U}_R(t, m, a)\,.$$

Here is a simple example to illustrate this construction. Suppose we are interested in explaining the simple scalar inference from (8) to (10). We consider two alternative forms $M = \{m_{\text{some}}, m_{\text{all}}\}$ where, obviously, choice of m_{some} represents an utterance of the sentence (8), and m_{all} corresponds to (9). This allows us to distinguish two states within the denotation of the target message m_{some}: in state $t_{\exists\neg\forall}$ m_{some} is true, while m_{all} is false; in state $t_{\forall}$ both messages are true. Together we obtain the interpretation game in Figure 2. (The table lists the prior probabilities for each state, the payoffs for each state-interpretation pair for S and R respectively, and the semantic meaning of messages. Most of this information is redundant when it is clear that we deal with interpretation games.)

A slightly more complex example is the context model in Figure 3 which we construct for the basic free choice reading of a sentence like (15a), with alternatives as in (20a) and (20b).

[23] This is simplified, and we will come back to the sender's utilities later when we also want to factor in different costs that different messages may have.

	$\Pr(t)$	$t_{\exists\neg\forall}$	$t_{\forall}$	m_{some}	m_{all}
$t_{\exists\neg\forall}$	$1/2$	1,1	0,0	$\surd$	$-$
$t_{\forall}$	$1/2$	0,0	1,1	$\surd$	$\surd$

Figure 2: Interpretation game for "some A's are B's"

	$\Pr(t)$	t_{A}	t_{B}	t_{AB}	$m_{\Diamond\text{A}}$	$m_{\Diamond\text{B}}$	$m_{\Diamond(\text{A}\vee\text{B})}$
t_{A}	$1/3$	1,1	0,0	0,0	$\surd$	$-$	$\surd$
t_{B}	$1/3$	0,0	1,1	0,0	$-$	$\surd$	$\surd$
t_{AB}	$1/3$	0,0	0,0	1,1	$\surd$	$\surd$	$\surd$

Figure 3: Interpretation game for "$\Diamond(A \vee B)$"

6.2 Epistemic Interpretation Games

Signaling games standardly incorporate the assumption that the sender knows the actual state. For base-level interpretation games, this means that these context models have a strong speaker competence assumption already built in: the interpretation behavior we derive in these context models serves to account for factual implicatures unmitigated by considerations of the sender's beliefs and opinions.

In order to extend quantity reasoning to epistemic implicatures we also need to accommodate for the possibility of epistemic uncertainty of the sender. The most conservative way of doing so is to stick to the signaling game framework and to simply "epistemicize" the notion of a state: epistemic interpretation games feature a set of states that distinguishes different epistemic states of the speaker, who then still is perfectly knowledgable about her own state of mind. The hearer's response actions are interpretations as to which epistemic state the speaker is in. So, which epistemic states of the speaker should the hearer distinguish?

The idea is the exact same as for base-level games. Epistemic states that are relevant to quantity reasoning about the meaning of S given Alt are obtained by looking at all (non-nontrivial) epistemic states in which the target sentence S is *believed to be true*. In that set, we lump together all those epistemic states which cannot be distinguished by different "*belief-value*s" the speaker may have with regard to different elements in Alt. Recall that in the classical framework that we work in, there are three belief values for any given proposition: (i) the agent believes the proposition is true (belief value "1"), (ii) the agent believes it is false (belief value "0"), or (iii) the agent is uncertain

about it (belief value "u"). Let's therefore say that two epistemic states X and Y —sets of possible worlds— are ALT-indistinguishable, $X \sim_{\text{ALT}} Y$, iff for all $A \in \text{ALT}$:

1. $X \subseteq A \Leftrightarrow Y \subseteq A$, and
2. $X \cap A \neq \emptyset \Leftrightarrow Y \cap A \neq \emptyset$.

The set of state distinctions for an epistemic interpretation game for S given ALT is then given as:

$$T_{\text{ES}} = \{\{X \subseteq S \mid X \sim_{\text{ALT}} Y\} \mid Y \subseteq S \wedge Y \neq \emptyset\} .$$

Accordingly, the semantic denotation function in epistemic interpretation games should be interpreted as: $t \in [\![m]\!]$ iff $\bigcup t \subseteq A$ for the unique alternative A that corresponds to m. — The rest stays, more or less, the same.

Here is a simple example. Suppose we are interested in the epistemic quantity implicatures standardly associated with the simple scalar contrast between "some" and "all", as in (8). The base-level interpretation game for this is given in Figure 2 and has two state distinctions $t_{\exists\neg\forall}$ and $t_{\forall}$. The set of epistemic states within those where the sender believes that "some" is true that we can distinguish based on this are:[24]

- the epistemic state $t_{[\forall]} = \bigcup\{t_{\forall}\}$ in which the sender believes that "all" is true;
- the epistemic state $t_{[\exists\neg\forall,\, \forall]} = \bigcup\{t_{\exists\neg\forall}, t_{\forall}\}$ in which the sender is uncertain whether "all" is true;
- the epistemic state $t_{[\exists\neg\forall]} = \bigcup\{t_{\exists\neg\forall}\}$ in which the sender believes that "all" is false.

This gives rise to the epistemic interpretation game in Figure 4.

Epistemic interpretation games can also incorporate various assumptions of the hearer about the speaker's competence. A plausible locus for implementing these assumptions are the prior probabilities. For instance, suppose that priors in the game from above are flat: $a = b$. This would encode the hearer's total uncertainty whether the speaker is in one epistemic state or another. On the other hand, if we wanted to implement the hearer's assumption that the speaker is competent in the model in Figure 4, we could do so by

[24] The set T_{ES} can often be alternatively depicted as $T_{\text{ES}} = \{\bigcup X \mid X \in \mathcal{P}(T_{\text{BL}}) \wedge X \neq \emptyset\}$ where T_{BL} is the corresponding set of base-level state distinctions. (Whether it can or cannot, depends on properties of ALT.) As in this example, I will therefore frequently write $t_{[x,y]}$ for an epistemic state in which the speaker considers only the base-level states t_x and t_y possible.

	$\Pr(t)$	$t_{[\exists\neg\forall]}$	$t_{[\forall]}$	$t_{[\exists\neg\forall,\forall]}$	m_{some}	m_{all}
$t_{[\exists\neg\forall]}$	a	1,1	0,0	0,0	$\surd$	$-$
$t_{[\forall]}$	a	0,0	1,1	0,0	$\surd$	$\surd$
$t_{[\exists\neg\forall,\forall]}$	b	0,0	0,0	1,1	$\surd$	$-$

Figure 4: Epistemic interpretation game for scalar implicature "some/all"

assuming that $a > b$. This would implement the idea that the hearer considers it *less likely* that the speaker is uncertain, but that he has otherwise no interpretation-relevant beliefs about the speaker's beliefs.

This can be generalized as follows. If the hearer assumes the speaker to be competent, then states with less uncertainty are considered *a priori* more likely than states with more uncertainty. More concretely even, the assumption of sender competence takes the following form in the present framework:

(37) *Competence Assumption:* If the speaker is (believed) competent, then the prior probability $\Pr(t)$ of information state t is given by a strictly *decreasing* function of the number alternatives that t is undecided about, i.e., assigns the belief value "u".

This, in fact, still leaves some wiggle room, as to how much difference we allow in probability between states where the speaker is differently competent. We will come back to this issue in Section 8.3.

Of course, there is a third option, which is though often overlooked. As argued in Section 2, the hearer may also assume that the speaker is *not* competent. In that case, we may adopt the following incompetence assumption:

(38) *Incompetence Assumption:* If the speaker is (believed) incompetent, then the prior probability $\Pr(t)$ of information states t are given by a strictly *increasing* function of the number alternatives that t is undecided about, i.e., assigns the belief value "u".

This completes this brief exposition of context models. The question we should turn to next, is how to "solve" these interpretation games.

7 Strategies, Equilibria and Explanations

Games like those in Figures 2, 3, and 4 fix certain parameters of the utterance context —arguably all and only those that are relevant to quantity reasoning— and as such puts certain constraints on what players of this game could possibly belief about each other. But the game model does not fully determine

the players' beliefs and action choices either. Still, we saw in Section 5 that it is important to a rationalization of quantity implicatures to give a detailed specification of what, for instance, the speaker believes the hearer's interpretation of a given sentence will be. Towards this end, we should start by fixing a notion of a player's behavior in a game.

Behavior. A player's behavior in a game is captured in the concept of a *strategy*. A *pure sender strategy* is a function $s \in M^T$ and a *pure receiver strategy* is a function $r \in A^M$. Pure strategies define how a player behaves in each possible information state that she might find herself in during the game. As the sender knows the actual state, she can choose a message conditional on the state she is in. As the receiver does not know the actual state, but only the sent message, he can condition his choice of action only on the message that he observed. We say that a pair $\langle s, r \rangle$ of pure sender and receiver strategies is a *pure strategy profile*.

For example, the game in Figure 2 allows for a number of pure strategies. Obviously, the most natural one is:

$$s = \left\{ \begin{array}{lcl} t_{\exists\neg\forall} & \mapsto & m_{\text{some}} \\ t_{\forall} & \mapsto & m_{\text{all}} \end{array} \right\} \qquad r = \left\{ \begin{array}{lcl} m_{\text{some}} & \mapsto & t_{\exists\neg\forall} \\ m_{\text{all}} & \mapsto & t_{\forall} \end{array} \right\} \tag{7.1}$$

whose receiver part captures drawing the attested quantity inference that the use of m_{some} conveys that the actual state is $t_{\exists\neg\forall}$, and whose sender part also conforms to our intuition about speakers who conform to the Speaker Quantity Principle in (3). This is not the only pure strategy profile in this game, but this is the one that we would like to be selected by a suitable *solution concept*, that specifies what counts as good or optimal behavior, so as to count as an explanation of a quantity implicature.

Similarly, for the interpretation game in Figure 3, we could say that we had "explained" the attested quantity implicature if by some independently motivated criterion we could select the pure strategy profile:

$$s = \left\{ \begin{array}{lcl} t_{\text{A}} & \mapsto & m_{\Diamond\text{A}} \\ t_{\text{B}} & \mapsto & m_{\Diamond\text{B}} \\ t_{\text{AB}} & \mapsto & m_{\Diamond(\text{A}\vee\text{B})} \end{array} \right\} \quad r = \left\{ \begin{array}{lcl} m_{\Diamond\text{A}} & \mapsto & t_{\text{A}} \\ m_{\Diamond\text{B}} & \mapsto & t_{\text{B}} \\ m_{\Diamond(\text{A}\vee\text{B})} & \mapsto & t_{\text{AB}} \end{array} \right\}. \tag{7.2}$$

Nash Equilibrium. The most widely-known solution concept for games is that of a *Nash equilibrium*. The idea behind this solution is that an equilibrium characterizes a *steady state*, i.e., a strategy profile in which no player would strictly benefit if he deviated from that profile given that everybody else would conform. For our signaling games, a pure strategy profile $\langle s, r \rangle$ is a pure Nash equilibrium whenever for all $t \in T$ we have:

(i) $\mathrm{U}_S(t, s(t), r(s(t))) \geq \mathrm{U}_S(t, s'(t), r(s'(t)))$ for all $s' \in M^T$, and

(ii) $\mathrm{U}_R(t, s(t), r(s(t))) \geq \mathrm{U}_R(t, s(t), r'(s(t)))$ for all $r' \in A^M$.

Both of the intuitive strategy profiles in (7.1) and (7.2) are Nash equilibria, as is easy to check. But, unfortunately, they are not uniquely so. For instance, a profile which is like (7.1) but which reverses messages:

$$s = \left\{ \begin{array}{lcl} t_{\exists\neg\forall} & \mapsto & m_{\text{all}} \\ t_{\forall} & \mapsto & m_{\text{some}} \end{array} \right\} \qquad r = \left\{ \begin{array}{lcl} m_{\text{some}} & \mapsto & t_{\forall} \\ m_{\text{all}} & \mapsto & t_{\exists\neg\forall} \end{array} \right\}$$

is also a Nash equilibrium of the game in Figure 2, in which, moreover, both players are just as well off as in the intuitive equilibrium in (7.1). This is of course weird and we should therefore perhaps rule out that messages may be used untruthfully (after all, message m_{all} is not true in state $t_{\exists\neg\forall}$).

But even that will not select the desired profiles uniquely for all cases. For consider the following Nash equilibrium of the game in Figure 3:

$$s = \left\{ \begin{array}{lcl} t_{\text{A}} & \mapsto & m_{\Diamond(\text{A}\vee\text{B})} \\ t_{\text{B}} & \mapsto & m_{\Diamond\text{B}} \\ t_{\text{AB}} & \mapsto & m_{\Diamond\text{A}} \end{array} \right\} \quad r = \left\{ \begin{array}{lcl} m_{\Diamond\text{A}} & \mapsto & t_{\text{AB}} \\ m_{\Diamond\text{B}} & \mapsto & t_{\text{B}} \\ m_{\Diamond(\text{A}\vee\text{B})} & \mapsto & t_{\text{A}} \end{array} \right\}. \tag{7.3}$$

This is surely counterintuitive, but it is nonetheless a proper Nash equilibrium in which players achieve perfect communication at maximal payoffs throughout. This example is particularly worrying because there seems to be no obvious refinement of the equilibrium notion that rules out (7.3) in favor of (7.2). For instance, Parikh (2001) proposes to filter Nash equilibria by a secondary criterion of *Pareto-optimality*: a strategy profile is Pareto-optimal if any deviation to some player's benefit would be to some other player's detriment. The Nash equilibrium in (7.3) *is* Pareto-optimal in this sense. Another possible set of refinements is suggested by van Rooij (2008), namely the *Neo-Logism Proofness* criterion of Farrell (1993) and the *intuitive criterion* of Cho and Kreps (1987). Neither of these refinements has any bite on the equilibrium in (7.3), because there are no "surprise messages" (see below) to which both criteria could apply.[25]

The source of the problem is, arguably, an improper treatment of conventional semantic meaning in pragmatic reasoning (cf. Franke 2009, 2010). Even if we were to assume that speakers cannot speak untruthfully, or are severely punished when they do, traditional solutions are still too weak.[26] What we

[25]The exact same problem of motivating the "obvious" choice of profile (7.2) over (7.3) also haunts Geurts's more informal but very related account of free choice implicatures (see Geurts 2010, Chapter 6).

[26]More strongly even, it would be methodologically short-sighted, if not plainly wrong to make either assumption, because, as is established wisdom in game theory (Rabin 1990; Farrell 1993;

need is a general solution concept that properly accounts for the role of semantic meaning, *and* that selects uniformly for the pragmatically "correct solutions". This is what the solution concept does that the following section will introduce.

8 Iterated Best Response Reasoning

Nash equilibrium, even with traditional refinements thereof, is not only insufficient for our present pragmatic purposes but it is also not optimally suited to explaining empirical data of actual human choice making in strategic contexts. The *iterated best response model* (IBR model, for short) that this section introduces as an alternative pragmatic solution concept draws on these empirical findings, so I start with a brief overview of its empirical motivation.

Level-k Thinking. Recent years have witnessed a strong increase in interest in experimental approaches to strategic reasoning (Camerer 2003). Experimental results strongly suggest that equilibrium solution concepts are, though theoretically appealing, *not* the best predictors of human performance. Rather, there is ample empirical evidence for what has been called "level-k reasoning", i.e., best response reasoning over discrete steps to only a certain depth k (e.g. Ho et al. 1998; Crawford 2007; Crawford and Iriberri 2007). Intuitively speaking, such "level-k reasoning" proceeds from some psychologically salient strategy, then first considers a best response to that initially salient strategy, then a best response to that, and so on. In general, human reasoning often seems to be *overconfident* in such iterated best response sequences in the sense that it underestimate others (believing that the opponent reasons only up to level $k-1$) and thus overestimating the own ability (believing to play well with k levels of reasoning). However, on repeated trials agents can learn to apply higher levels of best response reasoning (cf. Camerer 2003, chapters 5–6).

Such "level-k reasoning" plausibly solves the above problem of equilibrium selection. The idea is to treat the semantic meaning of expressions as a focal, but crucially non-binding attractor of attention, i.e., as an initially *salient strategy* from which iterated best response reasoning departs. Unsophisticated level-0 players only stick to the semantic content of expressions: level-0 senders are believed to arbitrarily say something true, and level-0 receivers are believed to interpret a message literally, i.e., as if it was a true observation revealed by nature, not a strategic choice of a communicating agent. On higher

Farrell and Rabin 1996), which aspects of conventional meaning are *credible* is subject to strategic considerations and depends crucially on the degree of conflict of interests between speakers and hearers.

levels of sophistication, level-k players choose a best response to the belief that their opponent is a level-$(k-1)$ player. This process may lead to a fixed point and it is this fixed point behavior which explains, I shall demonstrate, the pragmatic enrichments of messages in interpretation games.[27]

In the remainder of this section I spell out two versions of a simple IBR model that are provably equivalent. The first one, the "heavy system" spelled out in Section 8.1, is formulated explicitly in terms of the players' probabilistic beliefs and rationality. This affords rather complex definitions in terms of probability distributions which not every working linguist may be familiar with. Therefore, I will also introduce a "light system" in Section 8.3 that dispenses with most of the formal machinery, and that operationalizes the workings of the "heavy system" in terms that are more familiar to a larger linguistic audience, and that are easier to apply when checking the model's predictions.

8.1 The Heavy System

IBR models are models of step-by-step reasoning. We will therefore define inductively differently sophisticated *player types*. A player type is defined here as a set of pure strategies: those pure strategies that a player of that level of sophistication may be expected to play. These player types are singled out by particular assumptions about the *epistemic states* of players. Definitions of the players' types for the relatively simple IBR model that I introduce here are motivated by three assumptions.[28] Firstly, two assumptions about the hierarchy of player types, namely that it is common belief among players that:

BASE: level-0 players are entirely unstrategic; and that

STEP: level-$(k+1)$ players act rationally based on an unbiased belief —to be defined below— that their opponent is a level-k player.

Secondly, an assumption regulating the impact of conventional meaning on players' belief formation and choices, the so-called *truth ceteris paribus assump-*

[27]This does *not* mean that I suggest that every implicature is established by explicitly calculating through such a sequence of level-k thinking. This would be ludicrous. Just as exhaustivity-based approaches are not meant to be —or so I hope— psychologically realistic descriptions of the actual processes of natural language interpretation, so is the model I propose here. The IBR model wants to describe a generalized and idealized pragmatic competence, but it does so in explicit Gricean, rationalistic terms.

[28]For clarity: this is not the formal claim that the player types, as I introduce them here, can be *derived* from these assumptions in a suitable epistemic model. (I do not know whether this is the case.) Rather these epistemic constraints serve as the *motivation* for the following definitions.

tion (TCP assumption) namely that it is common belief among players that:[29]

TCP: everybody will stick to the conventional semantic meaning if otherwise indifferent.

Assumptions BASE and TCP yield level-0 players whose behavior is unconstrained except for truthfulness and literal uptake. So, as inductive base, define S_0 as the set of all pure strategies that send a true message, and R_0 is the set of all pure strategies that interpret a message as true:[30]

$$S_0 = \left\{ s \in M^T \mid \forall t \in T : \ t \in [\![s(t)]\!] \right\}$$
$$R_0 = \left\{ r \in A^M \mid \forall m \in M : \ r(m) \in [\![m]\!] \right\} .$$

The inductive step needs a little more elaboration. If S_k and R_k are sets of pure sender and receiver strategies respectively, the sets S_{k+1} and R_{k+1} are defined as playing *rationally* given some belief that the opponent plays a strategy in S_k or R_k. This could in principle be implemented in many different ways. Most importantly, we could adopt many *prima facie* plausible constraints on the belief formation process of level-$(k+1)$ players. To keep the system simple, I will assume here that players form *unbiased beliefs*:[31] a level-$(k+1)$ sender (receiver) believes that any strategy in R_k (S_k) is *equally likely*. Level-$(k+1)$ behavior is defined as rational under this belief.

Towards a definition of rational behavior given a certain belief, let's focus on the sender side first. If, for instance, a level-$(k+1)$ sender has an unbiased belief in a given set R_k of pure strategies, then we can derive from that the sender's so-called *behavioral beliefs* —formally a function in $(\Delta(A))^M$— that

[29]This assumption is stronger than actually needed. It merely assures semantics-conform behavior after so-called "surprise messages" and when interpretations are "uninducible" (see Section 8.2).

[30]Strictly speaking, this is only well-defined if the underlying signaling game satisfies the conditions:

1. $[\![m]\!] \neq \emptyset$ for all m;
2. $[\![t]\!]^{-1} \neq \emptyset$ for all t.

Interpretation games, if constructed as described in Section 6, satisfy these conditions, whenever ALT contains no contradictions.

[31]Different systems of "level-k reasoning" differ exactly in this design choice. The *cognitive hierarchy model* of Camerer et al. (2004), for example, assumes that players of level $k+1$ form the belief that their opponent is of level $l \leq k$. This may seem more realistic, but strongly complicates the mathematics. Other viable options are implemented by Jäger and Ebert (2009) or Mühlenbernd (2009). For our present purposes, the assumption of unbiased beliefs is welcome, chiefly because it helps simplify the system tremendously. But the assumption of unbiased beliefs also does some good pragmatic work for us: it implements a particular form of *forward induction reasoning* without which weaker systems, such as that by Jäger and Ebert (2009), cannot derive, for instance, the free choice implicatures of (15a) (see Franke 2009, for discussion and comparison).

represents how likely the sender believes it is that a given action is played in response to a given message. For unbiased beliefs in R_k this is not difficult to compute: the probabilistic sender belief that m is answered by a is simply the proportion of receiver strategies in R_k that map m to a. Abusing notation, we write:

$$R_k(m,a) = \frac{|\{r \in A^M \mid r(m) = a\}|}{|R_k|}.$$

Notice that since the only thing that the sender does not know when it comes to her making a move is the receiver's behavior, the sender's game relevant uncertainty is entirely captured by a behavioral receiver strategy. Using the standard definition of rationality as maximization of expected utility, we define the set of all *best responses* to the unbiased belief R_k as:

$$\text{BR}(R_k) = \left\{ s \in M^T \mid s(t) \in \arg\max_{m \in M} \sum_{a \in A} R_k(m,a) \times \text{U}_S(t,m,a) \right\}.$$

In line with the TCP assumptions from above, we then define the behavior of a level-$(k+1)$ sender as a best response in $\text{BR}(R_k)$ that, if possible, respects semantic meaning:

$$S_{k+1} = \left\{ s \in \text{BR}(R_k) \mid \forall t \ (\exists s' \in \text{BR}(R_k) \ t \in [\![s'(t)]\!]) \rightarrow t \in [\![s(t)]\!] \right\}.$$

A largely parallel definition yields higher level receiver types. However, since the receiver is uncertain about, not only what the sender does, but also about what state the sender has observed, the definition of the receiver's beliefs are slightly more complicated. As before, though, we can define the receiver's *behavioral beliefs* —formally: a function in $(\Delta(M))^T$— given an unbiased belief in S_k as:

$$S_k(t,m) = \frac{|\{s \in M^T \mid s(t) = m\}|}{|S_k|}.$$

These beliefs, however, only indirectly feed into the definition of receiver rational behavior, because it is primarily the receiver's *posterior beliefs* that decide what counts as a rational choice. The posterior beliefs $\mu \in (\Delta(T))^M$ specify how likely the receiver considers a state after observing a given message. Obviously, posterior beliefs should be a function of prior and behavioral beliefs wherever possible. The normatively correct way of forming posterior beliefs is by *Bayesian conditionalization*. We say that μ is *consistent* with Pr and S_k iff for all t and m for which there is a state t' such that $S_k(m|t') \neq 0$ we have:

$$\mu(t|m) = \frac{\Pr(t) \times S_k(m|t)}{\sum_{t' \in T} \Pr(t') \times S_k(m|t')}.$$

Consistency effectively demands conservative belief dynamics: wherever possible Bayesian conditionalization computes backward the *likelihood* for each state t that an observed message m was sent in t given t's prior probability and the probability with which m was expected to be sent in t.

With this we can define what a best response to a posterior belief μ is:

$$\mathrm{BR}(\mu) = \left\{ r \in A^M \mid r(m) \in \arg\max_{a \in A} \sum_{t \in T} \mu(t|m) \times \mathrm{U}_R(t, m, a) \right\} .$$

This gives us the proper definition of a best response to an unbiased belief in S_k:

$$\mathrm{BR}(S_k) = \{ \mathrm{BR}(\mu) \mid \mu \text{ is consistent with Pr and } S_k \} .$$

Under the TCP assumption, level-$(k+1)$ receiver behavior is then defined as:

$$R_{k+1} = \left\{ r \in \mathrm{BR}(S_k) \mid \forall m \ (\exists r' \in \mathrm{BR}(S_k) \ r'(m) \in [\![m]\!]) \rightarrow r(m) \in [\![m]\!] \right\} .$$

This is a long list of definitions. I will show presently in Section 8.3 that the behavior of this model can be expressed much more easily by reasoning about the cardinality of certain sets rather than about probabilistic beliefs and rationality. But before doing so, we should look at at least one example in the full-fledged system to understand better two essential concepts: (i) the effect of Bayesian conditionalization on pragmatic interpretation and (ii) the notions "surprise message" and "uninducible interpretation."

8.2 The Basic Free Choice Implicature in the Heavy System

To better understand the "heavy system" let us compute its predictions for the game in Figure 3, towards an explanation of the free-choice implicature of (15a). The unsophisticated receiver behavior in this game is given by the semantic meaning of messages only:[32]

$$R_0 = \left\{ \begin{array}{lcl} m_{\Diamond \mathrm{A}} & \mapsto & t_{\mathrm{A}}, t_{\mathrm{AB}} \\ m_{\Diamond \mathrm{B}} & \mapsto & t_{\mathrm{B}}, t_{\mathrm{AB}} \\ m_{\Diamond (\mathrm{A} \vee \mathrm{B})} & \mapsto & t_{\mathrm{A}}, t_{\mathrm{B}}, t_{\mathrm{AB}} \end{array} \right\} .$$

This defines the level-1 sender's unbiased behavioral belief: for example, S_1 believes that if she sends $m_{\Diamond \mathrm{A}}$ the receiver will not choose interpretation t_{B} at all, but may choose t_{A} or t_{AB} with equal probability. What is rational behavior under this belief? This can be calculated along the above definitions but it

[32] Sets of pure strategies $X \subseteq Z^Y$ are represented by listing for each $y \in Y$ the set of all $z \in Z$ such that for some strategy $x \in X$ we have $x(y) = z$.

is also intuitively appreciated that, for example, in state t_A the only rational choice given this belief is to send $m_{\diamond A}$: $m_{\diamond A}$ has a probability $1/2$ chance of inducing the correct response, while $m_{\diamond(A \vee B)}$ has a probability $1/3$ chance, and $m_{\diamond B}$ will simply never elicit the proper response in the hearer. Similar reasoning establishes:

$$S_1 = \left\{ \begin{array}{lcl} t_A & \mapsto & m_{\diamond A} \\ t_B & \mapsto & m_{\diamond B} \\ t_{AB} & \mapsto & m_{\diamond A}, m_{\diamond B} \end{array} \right\}.$$

It is noteworthy here that $m_{\diamond A}$ and $m_{\diamond B}$ are the best sender choices in t_{AB}, because under R_0's interpretation each of these messages yields a chance of $1/2$ of successful communication, as opposed to a chance of $1/3$ when sending $m_{\diamond(A \vee B)}$. That means that the target message $m_{\diamond(A \vee B)}$ will actually be a *surprise message* to R_2, as an unbiased belief in S_1 entails a belief that message $m_{\diamond(A \vee B)}$ never gets send. It is crucial to note here that Bayesian conditionalization —as it is needed to compute consistent posterior beliefs, and from there the proper responses of R_2— does not apply to surprise messages. There is indeed a lot of literature in rational choice theory dealing with how beliefs after surprise messages could or should be formed (cf. Stalnaker 1998). The present IBR model simply predicts that surprise messages could be answered by just any interpretation, were it not for the TCP assumption that, all else equal, players stick to the semantic meaning of messages. So, more or less naturally, we simply assume that surprise messages are interpreted *literally* like an unsophisticated receiver would. This gives us:

$$R_2 = \left\{ \begin{array}{lcl} m_{\diamond A} & \mapsto & t_A \\ m_{\diamond B} & \mapsto & t_B \\ m_{\diamond(A \vee B)} & \mapsto & t_A, t_B, t_{AB} \end{array} \right\}.$$

From this point on, the reasoning chain unfolds smoothly. Based on a belief in R_2, the sender will send message $m_{\diamond(A \vee B)}$ exactly in state t_{AB}, because in this state this is the only message that has a positive probability of inducing the right interpretation:

$$S_3 = \left\{ \begin{array}{lcl} t_A & \mapsto & m_{\diamond A} \\ t_B & \mapsto & m_{\diamond B} \\ t_{AB} & \mapsto & m_{\diamond(A \vee B)} \end{array} \right\}.$$

The only best response to this is for the receiver to interpret as follows:

$$R_4 = \left\{ \begin{array}{lcl} m_{\diamond A} & \mapsto & t_A \\ m_{\diamond B} & \mapsto & t_B \\ m_{\diamond(A \vee B)} & \mapsto & t_{AB} \end{array} \right\}.$$

This sequence of reasoning steps has thereby reached a fixed point. It is these fixed point of IBR reasoning which give the model's "idealized solution" in the sense that the behavior selected by any fixed point is compatible with what the most sophisticated agents in our model would play. (It is easy to see that fixed point behavior is rational behavior compatible with common belief in rationality.) In the present example, the fixed point then, in a way of speaking, *selects* the appropriate a Nash equilibrium that explains the free choice inference.

This is good, but we should also check the predictions of the IBR reasoning chain that starts with an unsophisticated sender.[33] It turns out that this reasoning chain indeed terminates in the exact same fixed point for mostly the same reasons. An unsophisticated sender is given by:

$$S_0 = \left\{ \begin{array}{lcl} t_{\mathrm{A}} & \mapsto & m_{\Diamond \mathrm{A}},\ m_{\Diamond(\mathrm{A} \vee \mathrm{B})} \\ t_{\mathrm{B}} & \mapsto & m_{\Diamond \mathrm{B}},\ m_{\Diamond(\mathrm{A} \vee \mathrm{B})} \\ t_{\mathrm{AB}} & \mapsto & m_{\Diamond \mathrm{A}},\ m_{\Diamond \mathrm{B}},\ m_{\Diamond(\mathrm{A} \vee \mathrm{B})} \end{array} \right\}.$$

In order to compute the set R_1 from this, we need to compute consistent posterior beliefs and then rational responses to these. Formally this is a mild load of work, but the rationale behind this reasoning can also be framed intuitively. For example, a level-1 receiver expects the message $m_{\Diamond \mathrm{A}}$ to be sent with probability $1/2$ in state t_{A}, with probability $1/3$ in state t_{AB} and not at all in state t_{B}. Consequently, by Bayesian conditionalization the state that the receiver thinks is *most likely* after hearing message $m_{\Diamond \mathrm{A}}$ is t_{A}. Since in an interpretation game with its particular payoff structure the rational choice is to go for the most likely interpretation, the receiver will therefore interpret $m_{\Diamond \mathrm{A}}$ as t_{A}. Similar reasoning leads us to verify that:

$$R_1 = \left\{ \begin{array}{lcl} m_{\Diamond \mathrm{A}} & \mapsto & t_{\mathrm{A}} \\ m_{\Diamond \mathrm{B}} & \mapsto & t_{\mathrm{B}} \\ m_{\Diamond(\mathrm{A} \vee \mathrm{B})} & \mapsto & t_{\mathrm{A}}, t_{\mathrm{B}} \end{array} \right\}.$$

Two observations are in order here. Firstly, the player types R_1 and S_1 look very similar and so feels the reasoning by which these were established. Although, strictly speaking, the structure of the receiver's beliefs and best responses is a fair deal more complicated than that of the sender's, it nearly seems as if we do not need to compute all posterior beliefs and best responses in detail after all — at least when dealing with interpretation games. This

[33] Other related approaches consider only consider pragmatic reasoning that starts from the assumption of a naïve listener (e.g. Stalnaker 2006; Benz and van Rooij 2007). Since I know of no good reason not to also assume that pragmatic reasoning might depart from the assumption of a naïve speaker too, I would like to stay on the safe side and check predictions of both approaches.

is indeed the case and the intuitive foundation of the "light system" to be spelled out in the next section.

Secondly, there is an analogue to "surprise messages" also on the sender side. If the sender has an unbiased belief in R_1 then she will believe that it is *impossible* to induce the interpretation t_{AB}: there simply is no message which, according to the sender's beliefs, would have the receiver select this interpretation. By the same reasoning as above the sender would then be indifferent between sending *any* message whatsoever, were it not, again, that we assume with TCP that the sender then *ceteris paribus* prefers to at least send a true message. This derives the player type:[34]

$$S_2 = \left\{ \begin{array}{lcl} t_{\mathrm{A}} & \mapsto & m_{\Diamond \mathrm{A}} \\ t_{\mathrm{B}} & \mapsto & m_{\Diamond \mathrm{B}} \\ t_{\mathrm{AB}} & \mapsto & m_{\Diamond \mathrm{A}},\ m_{\Diamond \mathrm{B}},\ m_{\Diamond (\mathrm{A} \vee \mathrm{B})} \end{array} \right\}.$$

The remaining steps of this IBR reasoning chain are straightforward. We reach the desired fixed point with the following two steps:

$$R_3 = \left\{ \begin{array}{lcl} m_{\Diamond \mathrm{A}} & \mapsto & t_{\mathrm{A}} \\ m_{\Diamond \mathrm{B}} & \mapsto & t_{\mathrm{B}} \\ m_{\Diamond (\mathrm{A} \vee \mathrm{B})} & \mapsto & t_{\mathrm{AB}} \end{array} \right\} \qquad S_4 = \left\{ \begin{array}{lcl} t_{\mathrm{A}} & \mapsto & m_{\Diamond \mathrm{A}} \\ t_{\mathrm{B}} & \mapsto & m_{\Diamond \mathrm{B}} \\ t_{\mathrm{AB}} & \mapsto & m_{\Diamond (\mathrm{A} \vee \mathrm{B})} \end{array} \right\}.$$

In conclusion, both strands of IBR reasoning single out the intuitive "pragmatically correct" behavior in the signaling game that models a standard context of utterance of a sentence like (15a). It is as if IBR reasoning selects uniquely a particular Nash equilibrium.

General Results. This raises the general questions whether IBR reasoning always terminates in a fixed point, and, if it does, whether this fixed point is an equilibrium solution of some variety? The answers to both of these questions are affirmative under certain provisos. Firstly, IBR reasoning always reaches a fixed point, if we are dealing with finite games where sender and receiver have aligned preferences (see Appendix B.1 for a proof):[35]

Theorem 8.1. For a signaling game of pure cooperation where $\mathrm{U}_S = \mathrm{U}_R$ and where M, A and T are finite, each IBR sequence reaches a fixed point.

[34] It transpires here that the TCP assumption is necessary for assuring truthful production when interpretations are uninducible, as well as literal interpretation of surprise messages. In fact, this is its only impact on behavior of IBR types with $k > 0$, and it establishes that *everywhere* in IBR reasoning about interpretation games, players adhere to the conventional meaning of signals (see Lemma B.5 and its proof in Appendix B.3).

[35] Notice that it is, however, not necessarily the case that each IBR reasoning sequence reaches the *same* fixed point. I currently know of no natural condition that would depict the class of games where both IBR reasoning strands necessarily converge. This issue, however, is also only of marginal relevance to pragmatic applications.

Secondly, in interpretation games, this fixed point will always be a *perfect Bayesian equilibrium*, a mild refinement of the above notion of Nash equilibrium (see Appendix B.2 for the necessary definitions and a proof):

Theorem 8.2. Any $\langle S^*, R^* \rangle$ that is a fixed point of an IBR sequence for a signaling game that satisfies conditions C1 and C2 gives rise to a perfect Bayesian equilibrium:

C1: $T = A$;

C2: $\mathrm{U}_{S,R}(t, m, a) = 1$ if $t = a$ and 0 if $t \neq a$.

The next section gives a handy algorithm which captures IBR reasoning in interpretation games. This will help answer other more general questions, such as: How does the system deal with epistemic quantity implicatures? How does the IBR model relate to exhaustive interpretation?

8.3 The Light System

While going through the example in the previous section, we had already observed that the computation of, say, R_1 and S_1 seemed strikingly similar, despite the asymmetry in accessible information. For instance, when computing the best response of a sender S_{k+1} in a state t, what we apparently have to do is the following. Firstly, we look at the set of messages that could in principle trigger the correct interpretation t; let's fix this set by:[36]

$$R_k^{-1}(t) = \{m \in M \mid \exists r \in R_k \ : \ r(m) = t\} \ .$$

Secondly, we need to ask whether this set is empty or not. If it is empty, then this means that t is not inducible, and according to the TCP assumption, the sender will choose any message that is true in t. If, on the other hand $R_k(t)$ contains at least one message which could trigger the right interpretation, then the best choice of the sender is that message in $R_k(t)$ which makes it *most likely* that t is chosen. Which one is this? Since we are dealing with unbiased beliefs, this will be the message in $R_k(t)$ for which the set

$$R_k(m) = \{t \in T \mid \exists r \in R_k \ : \ r(m) = t\}$$

is minimal.

Surprisingly enough, the same reasoning also applies to computing sophisticated receiver types (see the proof of Theorem 8.3 for details). Indeed

[36]Notice that I am again overloading notation. Here and in the following I take the liberty of identifying an arbitrary set of pure strategies $X \subseteq Z^Y$ by a function $f \in \mathcal{P}(Z)^Y$ such that $f(y) = \{z \in Z \mid \exists x \in X \ : \ x(y) = z\}$, and to address that function f also as X.

we can give an alternative definition of the previous IBR model as follows:

$$\begin{aligned}
\check{R}_0(m) &= [\![m]\!] \\
\check{S}_0(t) &= \check{R}_0(t)^{-1} \\
\check{S}_{k+1}(t) &= \begin{cases} \arg\min_{m \in \check{R}_k^{-1}(t)} |\check{R}_k(m)| & \text{if } \check{R}_k^{-1}(t) \neq \emptyset \\ \check{S}_0(t) & \text{otherwise.} \end{cases} \\
\check{R}_{k+1}(m) &= \begin{cases} \arg\min_{t \in \check{S}_k^{-1}(m)} |\check{S}_k(t)| & \text{if } \check{S}_k^{-1}(m) \neq \emptyset \\ \check{R}_0(m) & \text{otherwise.} \end{cases}
\end{aligned}$$

Notice that this definition is easier because (i) sender and receiver behavior is defined entirely in parallel and (ii) because we do not need to compute *any* probabilities — we only need to compare the cardinality of sets. Nonetheless, the above definition is equivalent to the "heavy system" in interpretation games with flat priors.

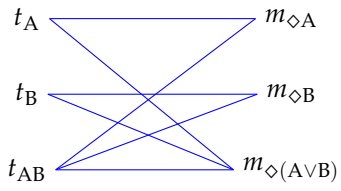
Theorem 8.3. The "light system" is an equivalent reformulation of the previous "heavy system", i.e., $R_k = \check{R}_k$ and $S_k = \check{S}_k$ for all k, if we assume that the game model satisfies conditions C1 and C2 from Theorem 8.2 plus the additional condition:

C3: $\Pr(t) = \Pr(t')$ for all t, t'.

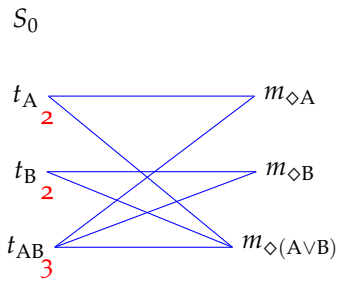
A proof is given in Appendix B.3.

This reformulation may not immediately seem that much simpler, but it actually is because it gives rise to a manageable algorithm with which the IBR reasoning can be computed with the help of simple diagrams. A set of pure strategies, such as for instance S_0 from the game in Figure 3, can be represented by the corresponding mappings of all $S_0(t)$ as follows (the mapping is obviously in left-to-right direction):

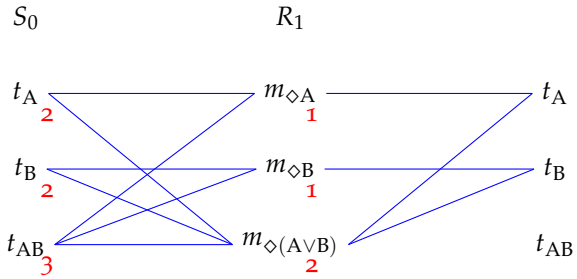
S_0



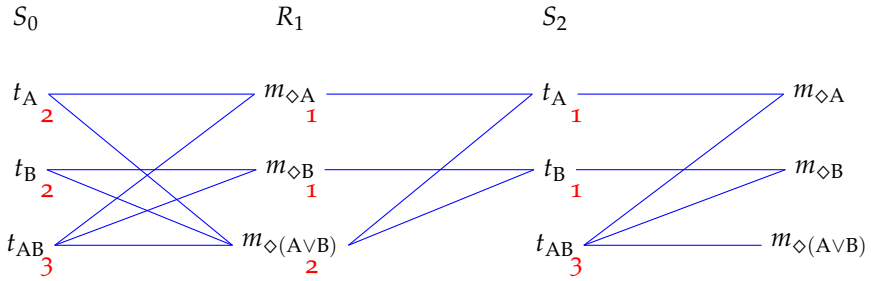
If we want to compute the best response of R_1 to this behavior, then, according to the light-weight reformulation, we simply need to count the number of outgoing connections from each state:



To plot our level-1 receiver we will then draw a connection from each message m to all those states that are connected with m that have the lowest number of outgoing connections among states connected with m:



Whenever, as in the computation of S_2, a node has no incoming connections, we restore the connections from the initial level-0 mapping:



One more step takes us to the fixed point. The full reasoning sequence of both IBR chains is represented in Figure 5.

Games with Non-Flat Priors. Theorem 8.3 gives us a simple algorithm to compute the predictions of the IBR model when priors are flat. This, then also applies to epistemic interpretation games in which we do not assume that the speaker is competent, such as when $a = b$ in the game in Figure 4. The result of running the IBR algorithm in this case is plotted in Figure 6. Both strands of IBR reasoning lead to the general epistemic implicature that the speaker does not believe that the alternative "all" is true.

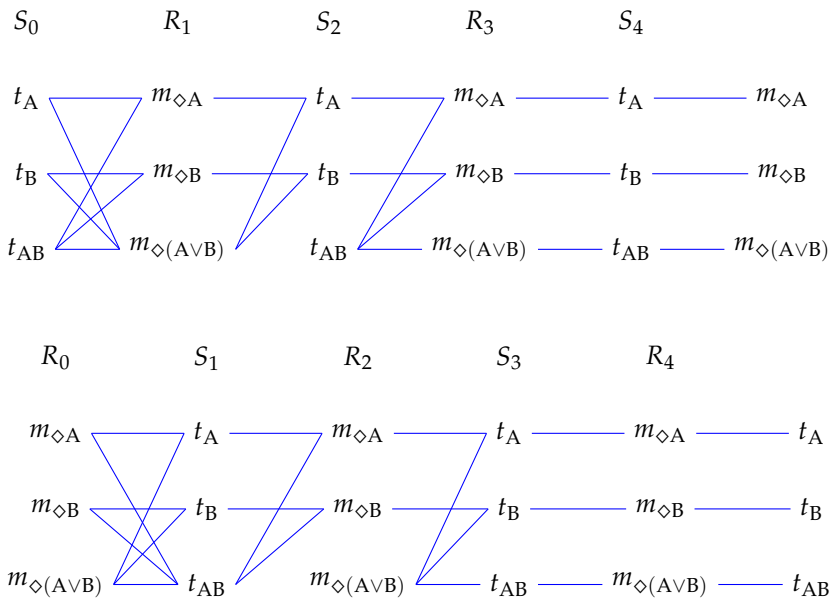


Figure 5: Schematic IBR reasoning for the game in Figure 3

Unfortunately, we cannot use the same algorithm to solve interpretation games with non-flat priors. Still, this does not necessarily mean that we always have to compute all probabilities and expected utilities in detail. If the differences between the prior probabilities are *small enough* we can treat them *as if* they were a second-order selection criterion on top of the interpretations selected under flat priors. This makes for a similarly easy algorithmic implementation of IBR reasoning in games with non-flat priors.

Theorem 8.4. Let $t_{\max}$ and $t_{\min}$ be the most and least likely states of a signaling game that satisfies conditions C1 and C2 of Theorem 8.1, plus

$$\frac{\Pr(t_{\min})}{\Pr(t_{\max})} < \frac{|M|-1}{|M|}\,.$$

Then for all $k \geq 0$:

$$R_{k+1}(m) = \left\{ t \in \breve{R}_{k+1}(m) \mid \neg\exists t' \in \breve{R}_{k+1}(m) \;:\; \Pr(t') > \Pr(t) \right\}.$$

Proof. We need to show that the given condition implies that the only case where prior probabilities ever make a difference between the posterior likelihood of two states t and t' given some message m is when $S_k(t,m) = S_k(t',m)$. To show this it suffices to look at $t_{\min}$ and $t_{\max}$ and the "worst case" where

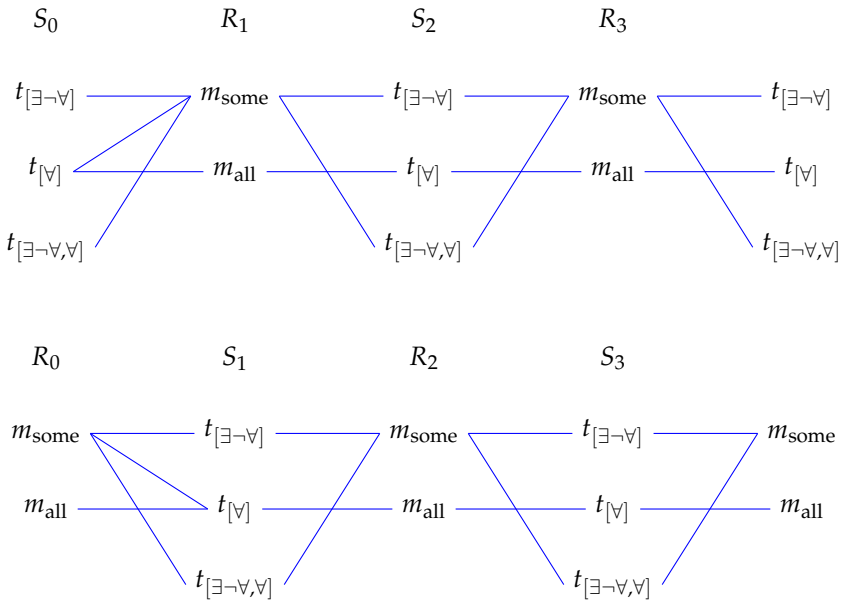


Figure 6: Predictions for the game in Figure 4 when $a = b$

$S_k(t_{\text{min}}, m) > S_k(t_{\text{max}}, m)$. We need to show that

$$\Pr(t_{\text{min}}) \times S_k(t_{\text{min}}, m) > \Pr(t_{\text{max}}) \times S_k(t_{\text{max}}, m)$$

even if the difference between $S_k(t_{\text{min}}, m)$ and $S_k(t_{\text{max}}, m)$ is as small as it can possibly get. Since we are dealing with interpretation games, this "worst case" is when $S_k(t_{\text{min}}, m) = \frac{1}{|M|-1}$ and $S_k(t_{\text{max}}, m) = \frac{1}{|M|}$. But if the priors satisfy the required condition, then the above holds true. □

This result tells us that when differences between priors are small enough we can calculate receiver behavior as follows: firstly we compute the receiver's possible interpretations $\breve{R}_k(m)$ as before; subsequently we choose from $\breve{R}_k(m)$ those states that maximize $\Pr(\cdot)$. That means that if differences in priors are small enough only the qualitative information in $\Pr(\cdot)$ is relevant, and it applies *as if* this was the secondary selection criterion in a lexicographic ordering after we have evaluated regular quantity reasoning.

For instance, if we look at the game in Figure 4 again, and if we adopt the competence assumption in (37) to the effect that a is *slightly* bigger than b, then we can calculate $R_1(m_{\text{some}})$ in three steps (see Figure 7 for the full results). Firstly, we look at all the states where m_{some} gets send:

$$[\![m_{\text{some}}]\!] = \left\{ t_{[\exists\neg\forall]}, t_{[\forall]}, t_{[\exists\neg\forall,\forall]} \right\} .$$

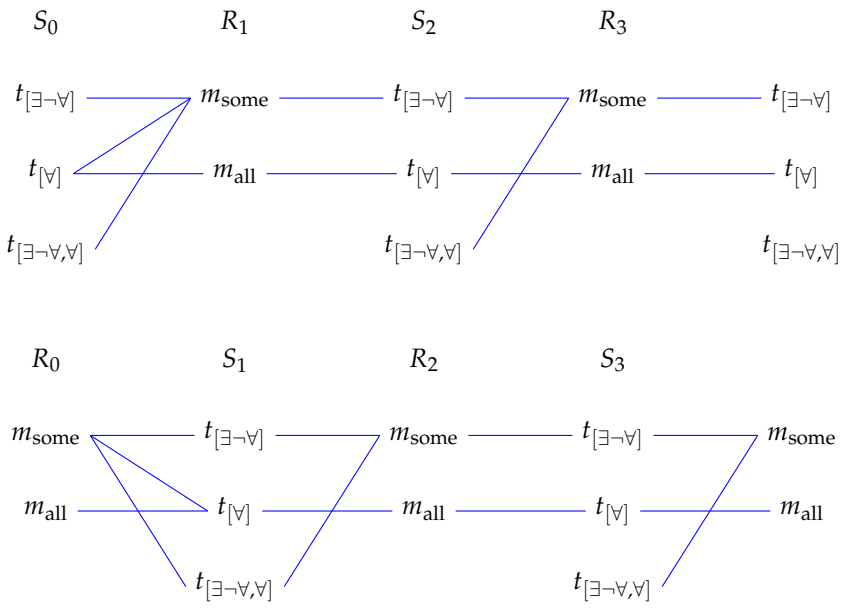


Figure 7: Predictions for the game in Figure 4 when $a > b$

Next, we select for the states in $[\![m_{\text{some}}]\!]$ in which fewest signals are chosen by S_0 — these are:

$$\left\{t_{[\exists\neg\forall]}, t_{[\exists\neg\forall,\forall]}\right\}.$$

Finally, we filter the *a priori* most likely states in this set — which leaves us with just $t_{[\exists\neg\forall]}$. The IBR model thus predicts that, under a sender competence assumption, the hearer comes to believe that a speaker who produces m_{some} believes that m_{all} is false.

Similarly, under the sender incompetence assumption in (38) we actually derive that m_{some} is associated with $t_{[\exists\neg\forall,\forall]}$. This is the inference that the speaker has no opinion as to whether the alternative m_{all} was true.

8.4 Relation to Exhaustive Interpretation

Using this graphical method of solving games with the IBR model will come in handy when checking predictions for several interesting cases in Section 9. But the reformulation into a "light system", and the algorithmic perspective that comes with it, also has a welcome theoretical spin-off. Look at the way R_1 is defined in the "light system": for any message m, R_1 selects those states in which m is true and in which fewest alternatives are true (from the set of

states where m is true):

$$R_1(m) = \left\{ t \in [\![m]\!] \mid \neg\exists t' \in [\![m]\!] \,:\, |R_0(t')^{-1}| < |R_0(t)^{-1}| \right\}|\,.$$

Isn't this what exhaustive interpretation in terms of minimal models —as defined in (4.1) and (4.2)— does too? Well, almost!

Consider the case of base-level interpretation games and $\textsc{Exh}_{\text{MM}}$ as defined in (4.1). The minimal worlds according to $\textsc{Exh}_{\text{MM}}$ are obtained by the partial order $<_{\textsc{Alt}}$ defined in terms of set inclusion of $A(w) = \{A \in \textsc{Alt} \mid w \in A\}$, whereas the minimal states —sets of worlds— that R_1 chooses are obtained by a total order defined in terms of $|A(t)|$ where $A(t) = \{A \in \textsc{Alt} \mid t \subseteq \textsc{Alt}\}$. It is obvious that all worlds in every state in $R_1(m)$ will be in $\textsc{Exh}_{\text{MM}}$:

Fact 8.5. Base-level exhaustive interpretation in terms of minimal models, $\textsc{Exh}_{\text{MM}}$, is entailed by level-1 receiver interpretation: $\bigcup R_1(m_{\text{S}}) \subseteq \textsc{Exh}_{\text{MM}}(S)$.

The reverse, however, is not necessarily the case: it may be the case that two worlds $w, v \in S$ are both minimal with respect to $<_{\textsc{Alt}}$, while w makes more alternatives true than v. In that case, w would not occur in any state in $R_1(m_{\text{S}})$, but it would be in $\textsc{Exh}_{\text{MM}}(S)$. However under a natural condition that the set of alternatives is in a sense "*homogeneous*", it is actually the case that exhaustive interpretation coincides with level-1 interpretation.[37] Analogous remarks apply to epistemic interpretation via $\textsc{Exh}^{\text{GE}}_{\text{MM}}$, and level-1 interpretation behavior in epistemic interpretation games without competence assumption.

Whether inclusion or identity, this result is actually quite remarkable either way. Contrary to the wide-spread conviction that exhaustive interpretation is licensed by something like the Speaker Quantity Principle, it actually *does not require* the assumption of a rational speaker to justify exhaustive interpretation. With some due provisos on the nature of Alt, the operator $\textsc{Exh}_{\text{MM}}$ captures rational interpretation based on nothing further than an understanding of conventional semantic meaning. More strongly even, it is not necessarily the case that $\textsc{Exh}_{\text{MM}}$ also coincides with R_2, which captures rational interpretation based on the assumption that the speaker behaves rationally and forthcoming (see, for example, Figure 5 where $R_1 \neq R_2$). In sum, the IBR

[37] We could spell out "homogeneity", for example, as the requirement that there be an isomorphism between any two maximal consistently-excludable sets in Max-CE$(S, \textsc{Alt})$ that preserves entailment relations. But it is not essential to the purpose of this article to spell out this condition —or any other sufficient condition on Alt— in more detail. It suffices to note that all examples discussed here indeed make it true that $\textsc{Exh}_{\text{MM}}(S) = \bigcup R_1(m_{\text{S}})$. — It should also be added that the fact that $<_{\textsc{Alt}}$ is a partial ordering does some good pragmatic work in the framework of Schulz and van Rooij (2006) when we consider more fine-grained semantic entities as in dynamic semantics, where we also keep track of discourse referents.

model here gives rise to a rather iconoclastic conclusion: some "*quantity*" implicatures follow from *quality* alone when we assume that hearer interprets rationally.[38]

9 Predictions & Results

This section is dedicated to checking some of the predictions of the IBR model. The first part of this section deals with the implicatures of the three types of disjunctive constructions that were discussed in Section 3. The later parts deal with various interesting other cases, some extensions and problems for the IBR model.

9.1 Disjunctions: Checking the Test Cases

We would like to check whether the IBR model can account for all of the implicatures standardly associated with: (i) *plain disjunctions* of the form $A \vee B$ as in (11), (ii) FC-*disjunctions* of the form $\Diamond(A \vee B)$ as in (15a) and (iii) SDA-*disjunctions* of the form $(A \vee B) > C$ as in (21) or (25).[39] In particular, we'd like to see two *factual implicatures*: the free-choice inference for $\Diamond(A \vee B)$ and the SDA inference for $(A \vee B) > C$. Moreover, as we observed in Section 3, all of these three constructions give rise to exactly one variety of *epistemic implicature*: a plain disjunction $A \vee B$ is associated with the *ignorance implicature* in (12) that the speaker is uncertain about either disjunct: $\mathrm{Uc}_S A \wedge \mathrm{Uc}_S B$; an FC-disjunction can gives rise to $\mathrm{Uc}_S \Diamond A \wedge \mathrm{Uc}_S \Diamond B$ as in (17), and an SDA-disjunction can give rise to $\mathrm{Uc}_S(A > C) \wedge \mathrm{Uc}_S(B > C)$ as in (36). Additionally, we should also check —just for sports, i.e., irrespective of whether we believe that this is the generally correct explanation or not— whether the IBR model can derive *exclusivity implicatures*, such as in (13) and (18), as quantity implicatures if we also take into account the respective conjunctive alternatives.

Factual Implicatures. We have already seen how the IBR model deals with the basic free-choice readings of sentences of the form $\Diamond(A \vee B)$ as a factual implicature in a base-level interpretation game. The base-level context model was given in Figure 3 and it had three state distinctions, repeated here:

[38] A different rationalization of epistemic exhaustive interpretation in terms of minimal models is given by de Jager and van Rooij (2007). This characterization, however, relies on a different set-up of the signaling game and further strong assumptions about the set ALT that are incompatible with some of the central examples we look at here.

[39] For expository reasons, it is helpful to first address the simple case with only two logically independent propositions A and B for all of these constructions. Sections 9.2 and 9.3 level these assumptions.

	$m_{\Diamond \mathrm{A}}$	$m_{\Diamond \mathrm{B}}$	$m_{\Diamond(\mathrm{A} \vee \mathrm{B})}$
t_{A}	$\surd$	$-$	$\surd$
t_{B}	$-$	$\surd$	$\surd$
t_{AB}	$\surd$	$\surd$	$\surd$

If we assume that a plain disjunction $A \vee B$ has alternatives A and B, then we derive three isomorphic state distinctions:

	m_{A}	m_{B}	$m_{\mathrm{A} \vee \mathrm{B}}$
t_{A}	$\surd$	$-$	$\surd$
t_{B}	$-$	$\surd$	$\surd$
t_{AB}	$\surd$	$\surd$	$\surd$

Interestingly, the same holds for conditionals of the form $(A \vee B) > C$ with alternatives $A > C$ and $B > C$ under the simple order-sensitive semantics given in Section 3:

	$m_{\mathrm{A}>\mathrm{C}}$	$m_{\mathrm{B}>\mathrm{C}}$	$m_{(\mathrm{A} \vee \mathrm{B})>\mathrm{C}}$
t_{A}	$\surd$	$-$	$\surd$
t_{B}	$-$	$\surd$	$\surd$
t_{AB}	$\surd$	$\surd$	$\surd$

That means that all three constructions give rise to exactly the same context models —at base-level but also for epistemic interpretation games (see below). Consequently, we only have to check predictions of the IBR model once, as these will be identical for all three cases.

Indeed, we have already done so. The calculation for the game from Figure 3 were graphically depicted in Figure 5. The exact same reasoning applies to the other two cases, provided we reinterpret the state names according to the tables above. This result is good and bad. It is good, because free-choice and SDA are treated exactly alike by the IBR model. This improves on the predictions of exhaustivity-based approaches. But, on the other hand, IBR also weirdly predicts that a plain disjunction $A \vee B$ —if interpreted in base-level interpretation games without conjunctive alternative— is associated with the conjunctive meaning that *both* A and B are true: also in this case, disjunction is pragmatically enriched to conjunction! This cannot be correct, can it? Does that mean that the IBR approach is mistaken all over?

No, it doesn't. A nuanced assessment is necessary. The problem obviously is that a particular and frequent use of plain disjunctions is at odds with some of the assumptions inherent in base-level interpretation games, most palpably: that the speaker is perfectly knowledgable. If we take the systems

prediction seriously, then it should tell us that disjunctions should *normally* be interpreted in epistemic context models, where predictions are indeed flawless (as I will show below). This intuition is probably also what underlies certain non-standard semantics of disjunctions as inherently modal elements (Zimmermann 2000; Geurts 2005).

Nonetheless, it should also be noted that there are cases of plain wide-scope disjunctions which *can* receive (some sort of) conjunctive reading. First of all, wide scope disjunctions such as (39) *can* give rise to a conjunctive reading.[40]

(39) You may take an apple or you may take a pear. $\Diamond A \vee \Diamond B$

Other disjunctive constructions also give rise to rather puzzling conjunctive readings as evidenced in (40) (cf. Culicover and Jackendoff 1997; Gómez-Txurruka 2002; Franke 2008).

(40) a. I want you to leave now, or I will start to cry.

b. It's good that you left Berlin, or you would never have finished your thesis.

c. That is enough points of order, or we will never get on to the items on the agenda. (from corpus: Europarl (en) (EU-EN))

The use of disjunction in (40) can roughly be paraphrased with "because otherwise", and it gives rise to a conjunctive reading that both disjuncts are true (modulo proper restriction of the modal in the second clause). For instance, (40a) is most naturally read as:

(41) I want you to leave, because if you do not I will start to cry.

Examples such as these are certainly puzzling, and require a careful analysis of the interaction of connectives with modal subordination (Roberts 1989). On the surface, however, they may suggest that not all disjunctions block a conjunctive reading entirely.

This, of course, is still a far shot away from a general theory explaining when and how disjunctions receive conjunctive readings. I have no such theory at offer here. As for the IBR model, the safest position to adopt is that base-level interpretation of a disjunction is a murky business to begin with. Unless a special reading can be coerced by some secondary pragmatic reasoning process —which the IBR model does not capture— the use of a disjunction is at odds with some of the assumptions encoded in base-level interpretation

[40]This has been stressed, *inter alia*, by Kamp (1978), Zimmermann (2000), Geurts (2005), Schulz (2005), and Simons (2005).

games. So maybe we can turn a vice into a virtue and say that the contestable base-level predictions of the IBR model are possibly an explanation for exactly why plain disjunctions are most naturally interpreted in a context which accommodates the speaker's uncertainty — to which we turn next.

Epistemic Implicatures. The epistemic states that we can distinguish based on a plain disjunction $A \vee B$ with alternatives A and B are the following six states:[41]

	m_{A}	m_{B}	$m_{\mathrm{A}\vee\mathrm{B}}$
$t_{[1,0,1]}$	$\surd$	$-$	$\surd$
$t_{[0,1,1]}$	$-$	$\surd$	$\surd$
$t_{[1,1,1]}$	$\surd$	$\surd$	$\surd$
$t_{[1,\mathrm{u},1]}$	$\surd$	$-$	$\surd$
$t_{[\mathrm{u},1,1]}$	$-$	$\surd$	$\surd$
$t_{[\mathrm{u},\mathrm{u},1]}$	$-$	$-$	$\surd$

It is clear that six exactly parallel patterns of belief value distributions arise for $\Diamond(A \vee B)$ and $(A \vee B) > C$. It suffices therefore to stick with the case of plain disjunctions.

In order to encode different competence assumptions into the context, we should parameterize the prior probabilities as follows (recall that the number of 'u'-s is relevant in assumptions (37) and (38)):

	$t_{[1,0,1]}$	$t_{[0,1,1]}$	$t_{[1,1,1]}$	$t_{[1,\mathrm{u},1]}$	$t_{[\mathrm{u},1,1]}$	$t_{[\mathrm{u},\mathrm{u},1]}$
$\Pr(\cdot)$	a	a	a	b	b	c

We'd like to check that for any choice of parameter a, b and c, the proper ignorance reading is derived. This means that we would like to find $t_{[\mathrm{u},\mathrm{u},1]}$ as the only interpretation assigned to $m_{\mathrm{A}\vee\mathrm{B}}$ in all fixed points. In this state, the speaker is uncertain about A and B, but she believes in $A \vee B$. For plain disjunction $A \vee B$, this amounts to the ignorance implicature $\mathrm{Uc}_S A \wedge \mathrm{Uc}_S B$. (For the other disjunctive constructions the same applies with due changes in interpretation of $t_{[\mathrm{u},\mathrm{u},1]}$: for instance, for $\Diamond(A \vee B)$ this state implies $\mathrm{Uc}_S \Diamond A \wedge \mathrm{Uc}_S \Diamond B$.)

This is indeed the only interpretation selected for the target message in all six constellations we would need to check. (Two IBR sequences for three different parameter sets.) Here is an informal argument why this is so. For instance, R_1 will interpret $m_{\mathrm{A}\vee\mathrm{B}}$ as $t_{[\mathrm{u},\mathrm{u},1]}$, and only as $t_{[\mathrm{u},\mathrm{u},1]}$, because this is

[41] The indices of states are sequences of belief values in the order that messages are given in the table.

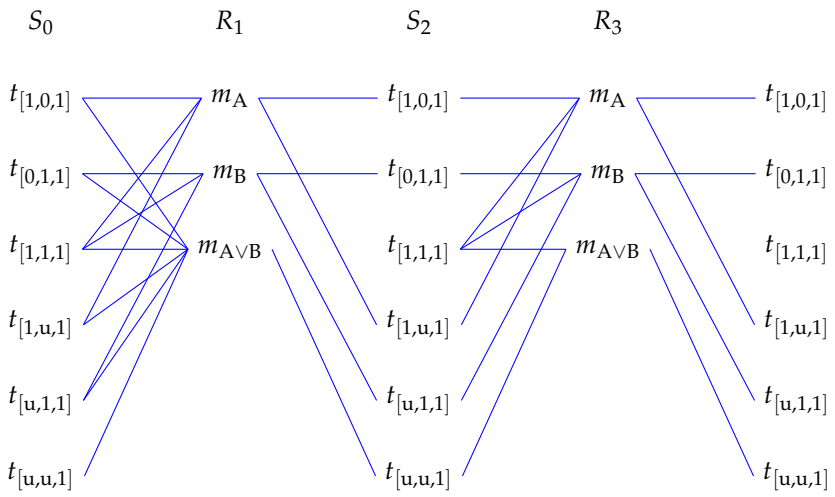


Figure 8: Predictions for epistemic interpretation of "A or B" when $a = b = c$

the state which minimizes the number of true messages in this game. For the same reason, S_1 will use $m_{\mathrm{A}\vee\mathrm{B}}$, and only $m_{\mathrm{A}\vee\mathrm{B}}$, in state $t_{[\mathrm{u},\mathrm{u},1]}$. Later steps of either sequence will not change this one-to-one association, and different parameters implementing different competence assumptions do not change this mapping either. (As one example of the six necessary calculations, the sequence starting with S_0 without competence assumption is given in Figure 8; the other calculations are boringly similar.)

Different competence assumptions do however influence the interpretation of messages other than $m_{\mathrm{A}\vee\mathrm{B}}$. In the absence of a competence assumption, the unique fixed point of IBR reasoning has m_{A} interpreted as either $t_{[1,0,1]}$ or as $t_{[1,\mathrm{u},1]}$ (see Figure 8). This captures the reading that the speaker believes that A is true, but does not believe that B is true. On the other hand, if we assume that the speaker is competent, we derive that m_{A} is interpreted as $t_{[1,0,1]}$: the implicature that the speaker knows that B is false. Lastly, under an incompetence assumption, we derive the interpretation $t_{[1,\mathrm{u},1]}$ for m_{A} which captures the reading that the speaker is uncertain about B. All of this accords neatly with intuition.

Exclusivity Implicatures at Base-Level. When we add a conjunctive alternative to the brew, the three types of disjunctive constructions we consider here no longer give rise to the same context models, so that we have to consult each case in turn. The contextual state distinction for base-level interpretation given $A \vee B$ with alternatives A, B and $A \wedge B$ are actually the same as before:

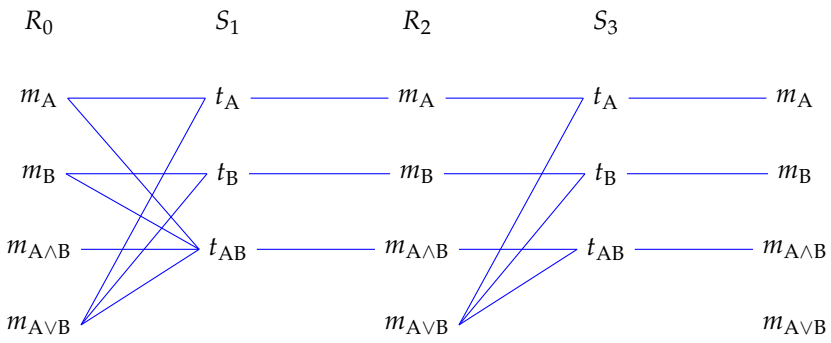


Figure 9: Predictions for plain disjunctions with conjunctive alternative

	m_{A}	m_{B}	$m_{\mathrm{A}\wedge\mathrm{B}}$	$m_{\mathrm{A}\vee\mathrm{B}}$
t_{A}	$\surd$	–	–	$\surd$
t_{B}	–	$\surd$	–	$\surd$
t_{AB}	$\surd$	$\surd$	$\surd$	$\surd$

In spite of this, the presence of an additional message will change the interpretation process. The R_0-part of the reasoning is summarized in Figure 9 where we see that the fixed point interpretation does not at all enrich the target disjunction. (The S_0-part is analogous.) Again the prediction is not surprising at all: if we construct a context representation based on the assumption that the speaker is totally knowledgable, then the use of a plain disjunction is a surprise to the hearer: in a context where information transmission counts, natural readings of disjunctions require to factor in the speaker's uncertainty.

For the target form $\Diamond(A \vee B)$ we derive a different set of state distinctions if we additionally consider the conjunctive alternative $\Diamond(A \wedge B)$, namely:

	$m_{\Diamond\mathrm{A}}$	$m_{\Diamond\mathrm{B}}$	$m_{\Diamond(\mathrm{A}\wedge\mathrm{B})}$	$m_{\Diamond(\mathrm{A}\vee\mathrm{B})}$
$t_{[1,0,0,1]}$	$\surd$	–	–	$\surd$
$t_{[0,1,0,1]}$	–	$\surd$	–	$\surd$
$t_{[1,1,1,1]}$	$\surd$	$\surd$	$\surd$	$\surd$
$t_{[1,1,0,1]}$	$\surd$	$\surd$	–	$\surd$

Unlike for plain disjunctions, a state $t_{[1,1,0,1]}$ is possible: it is possible for $\Diamond A$ and $\Diamond B$ to be true (there is an accessible A-world, and an accessible B-world), while $\Diamond(A \wedge B)$ is false (there is no accessible world in which both A and B are true).

The model's predictions for this case bare no surprises. Figure 10 shows the S_0-sequence. The fixed point interpretation of $\Diamond(A \vee B)$ establishes the

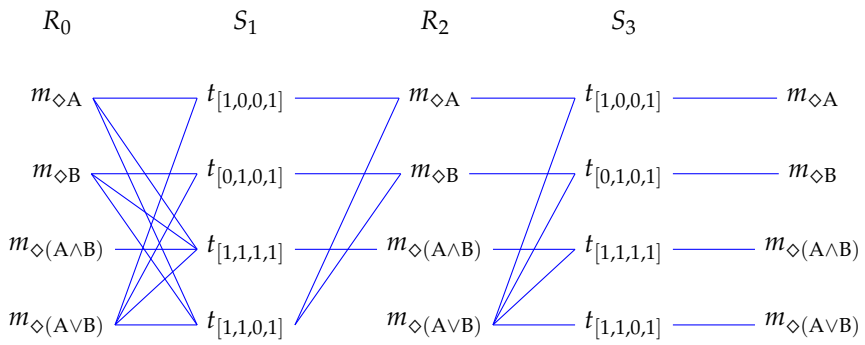


Figure 10: Predictions for FC-disjunctions with conjunctive alternative

free-choice reading, as well as the exclusivity implicature that $\Diamond(A \wedge B)$ is false.

Finally, the context model for factual interpretation of $(A \vee B) > C$ is even a little bigger. We can now distinguish six states, because it is also possible for either of $A > C$ or $B > C$ to be false when $(A \wedge B) > C$ is true. So, we get:

	$m_{A>C}$	$m_{B>C}$	$m_{(A\wedge B)>C}$	$m_{(A\vee B)>C}$
$t_{[1,0,0,1]}$	$\surd$	–	–	$\surd$
$t_{[0,1,0,1]}$	–	$\surd$	–	$\surd$
$t_{[1,0,1,1]}$	$\surd$	–	$\surd$	$\surd$
$t_{[0,1,1,1]}$	–	$\surd$	$\surd$	$\surd$
$t_{[1,1,0,1]}$	$\surd$	$\surd$	–	$\surd$
$t_{[1,1,1,1]}$	$\surd$	$\surd$	$\surd$	$\surd$

The predictions of the IBR model are summarized in Figure 11, where the R_0 sequence is spelled out. We see that the target message $(A \vee B) > C$ is interpreted correctly in the fixed point that is reached after R_4 to give rise to both the SDA inference, as well as the exclusivity implicature. The results for the S_0-sequence are identical for the interpretation of the target message.[42]

Exclusivity Implicatures at Epistemic Level. It remains to be checked what happens in *epistemic* interpretation games when we also have the conjunctive alternatives around. Although the context models differ and calculations are therefore not precisely the same, it turns out that the IBR model predicts notionally the exact same result for all of our three types of disjunctive constructions. I will therefore contend myself with discussing only the case of

[42]The S_0-sequence differs slightly only in the interpretation of $(A \wedge B) > C$.

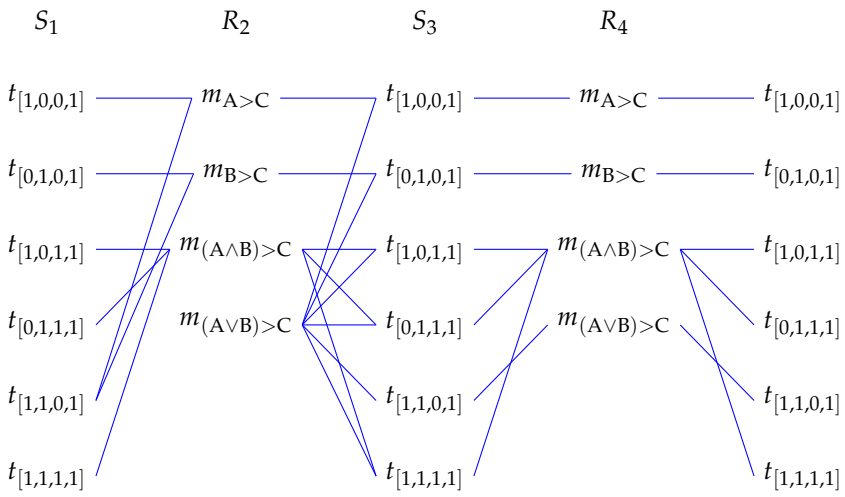


Figure 11: Predictions for SDA-disjunctions with conjunctive alternative

plain disjunction.

The interpretation $t_{[u,u,1]}$ selected for $m_{A\vee B}$ in the game in Figure 6 is uncommitted as to the belief value of the conjunction $A \wedge B$: $t_{[u,u,1]}$ is the set of epistemic states that are compatible with A, B and the negations thereof, and that contain only worlds where $A \vee B$ is true; this means that some epistemic states in $t_{[u,u,1]}$ will make $A \wedge B$ true, but others will not. But if we take $A \wedge B$ into account as an additional alternatives, we obtain a finer-grained set of state distinctions where different epistemic exclusivity implicatures can be expressed:

	m_A	m_B	$m_{A\wedge B}$	$m_{A\vee B}$
$t_{[1,0,0,1]}$	√	–	–	√
$t_{[0,1,0,1]}$	–	√	–	√
$t_{[1,1,1,1]}$	√	√	√	√
$t_{[u,u,0,1]}$	–	–	–	√
$t_{[1,u,u,1]}$	√	–	–	√
$t_{[u,1,u,1]}$	–	√	–	√
$t_{[u,u,u,1]}$	–	–	–	√

An example calculation for the sequence starting with S_0 in the absence of a competence assumption is given in Figure 12. (The other derivations are similar.) In this case we derive the general epistemic inference that the use of $A \vee B$ is associated with the set $\left\{t_{[u,u,0,1]}, t_{[u,u,u,1]}\right\}$ which captures that the

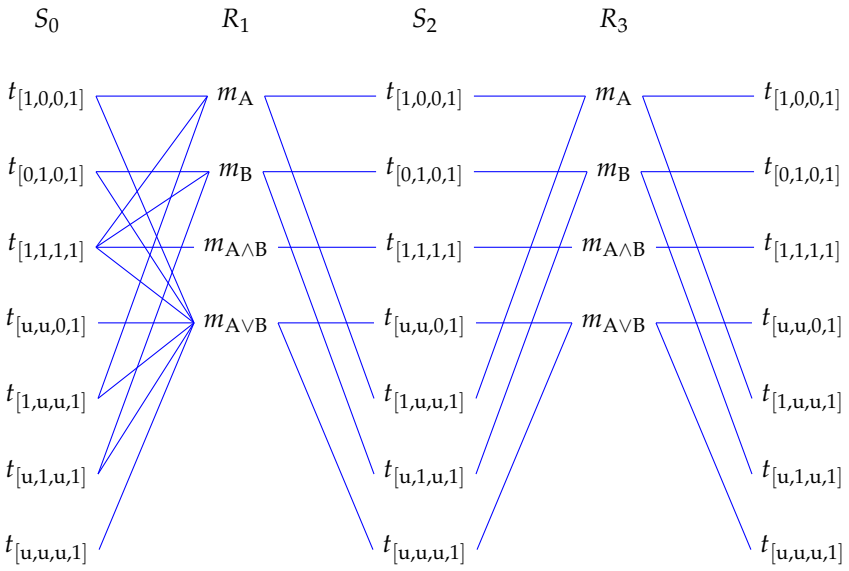


Figure 12: Predictions for the interpretation of "A or B" with conjunctive alternative and no competence assumption

speaker does not believe that $A \wedge B$ is true. If we integrate the competence assumption, we derive the implicature that $m_{\text{A}\vee\text{B}}$ is associated with the state $t_{[\text{u,u,0,1}]}$ only: the stronger implicature that the speaker knows that $A \wedge B$ is false. Finally, with an incompetence assumption, we derive that the target message is associated with $t_{[\text{u,u,u,1}]}$, which vindicates the implicature that the speaker is strictly uncertain about the truth of $A \wedge B$.

9.2 Reasoning with Message Costs

Entailing Disjuncts. So far we have assumed that disjuncts were logically independent. But not all disjunctions have this property, and it's there that we may expect problems with the current approach. For example, under truth-conditional semantics the sentence "A or (A and B)" is equivalent to the sentence "A." But, intuitively, these forms are to be interpreted differently, at least in certain contexts and if we assume that the speaker is competent on the issue at hand. Compare the answers (43a) and (44a) to a question (42).

(42) Who (of John and Mary) came to the party?

(43) a. John did.

b. $\rightsquigarrow$ The speaker knows that John came and that Mary did not.

(44) a. John or (John and Mary).

b. $\rightsquigarrow$ The speaker knows that John came and considers it possible that Mary came too.

Though equivalent in terms of truth-conditions, the implicatures associated with these answers, (43b) and (44b) respectively, are clearly different. This problem palpably affects pretty much all global Neo-Gricean accounts that rely on truth-conditional semantics, and it is thus interesting to see whether IBR can cope with this.

Of course, the most reasonable rejoinder here is to deny that truth-conditional semantics is suitable in the first place. The sentences (43a) and (44a) clearly do not have the same dynamic properties as evidenced in (45).

(45) a. ? John came to the party. The latter possibility is rather unlikely though.

b. John or (John and Mary) came to the party. The latter possibility is rather unlikely though.

This is why, for instance, Schulz and van Rooij (2006) use exhaustive interpretation in minimal *dynamic* models to account for cases of this kind. The IBR model could very well do the same thing, as it is not dependent of *any particular* notion of semantics, as long as meaning can be expressed in set-theoretic terms. There is, however, also another possibility, and that is to assume that the sentences A and $A \vee (A \wedge B)$ are, though equivalent, differently *costly*.

Suppose that for an interpretation of $A \vee (A \wedge B)$ we compare it simply to its disjuncts A and $A \wedge B$. This derives the following set of state distinctions for an epistemic interpretation game:

	m_{A}	m_{AB}	$m_{\mathrm{A}\vee(\mathrm{A}\wedge\mathrm{B})}$
$t_{[1,0,1]}$	$\surd$	$-$	$\surd$
$t_{[1,1,1]}$	$\surd$	$\surd$	$\surd$
$t_{[1,u,1]}$	$\surd$	$-$	$\surd$

Let us assume that message $m_{\mathrm{A}\vee(\mathrm{A}\wedge\mathrm{B})}$ is slightly more costly than its equivalent m_{A}. Let us also assume that these costs are *nominal*, as the economists would say, i.e., small enough that they apply —like prior probabilities in our IBR approach— as a *secondary selection criterion*. The IBR reasoning starting with R_0 under a competence assumption is spelled out in Figure 13. The crucial step in this computation is S_1 where message costs favor the sending of m_{A} in states $t_{[1,0,1]}$ and $t_{[1,u,1]}$ over sending $m_{\mathrm{A}\vee(\mathrm{A}\wedge\mathrm{B})}$.

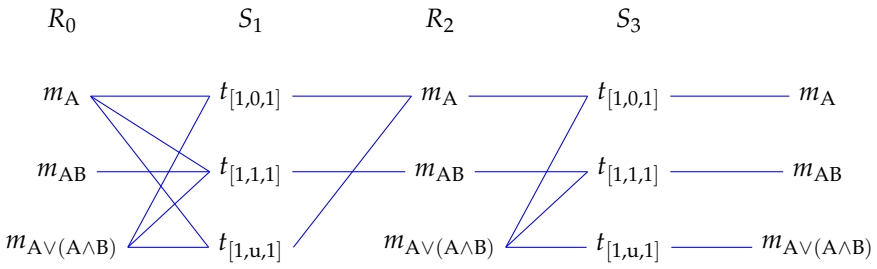


Figure 13: Predictions for interpretation of "A or (A and B)" under a competence assumption

Division of Pragmatic Labor. Another fairly systematic pattern of pragmatic inference that is squarely related to quantity implicatures is what has become known as *Horn's division of pragmatic labor* (cf. Horn 1984; Levinson 2000). It is a fairly ubiquitous phenomenon in natural languages that a simple way of expressing a meaning as in (46a) is associated with a stereotypical interpretation as in (46b), whereas a marked and overly complex way of expressing the very same meaning as in (47a) is interpreted in a non-stereotypical way as in (47b).

(46) a. Black Bart killed the sheriff.

b. $\rightsquigarrow$ Black Bart killed the sheriff in a stereotypical way.

(47) a. Black Bart caused the sheriff to die.

b. $\rightsquigarrow$ Black Bart killed the sheriff in a non-stereotypical way.

On closer look, Horn's division of pragmatic labor actually captures the interplay of two inferences. In abstract terms, there are two semantically equivalent expressions m and m_* both of which could denote either an unmarked state t, or a marked state t_*. Given that one expression m_* is more marked than the other m, the first part of the pragmatic inference pattern associates the unmarked form with the unmarked state of affairs ($m \rightsquigarrow t$); the second part of the pragmatic inference pattern associates the marked form with the marked state of affairs ($m_* \rightsquigarrow t_*$).

The IBR model can explain this inference pattern straightforwardly if we assume that (i) markedness of a meaning can be expressed in terms of a lower prior probability and that (ii) markedness of a form can be expressed as a higher message cost. With these assumptions, the base-level context model for this case is the one given in Figure 14 where ϵ and δ are sufficiently small values. The reasoning chains for this case are given in Figure 15. This is actually astonishing in the light of the difficulty that other game-theoretic

	$\Pr(\cdot)$	t	t_*	m	m_*
t	$.5+\epsilon$	1,1	0,0	$\surd$	$\surd$
t_*	$.5-\epsilon$	0,0	1,1	$\surd$	$\surd$
			$\text{cost}(\cdot)$	0	δ

Figure 14: Context model for Horn's "Division of Pragmatic Labor"

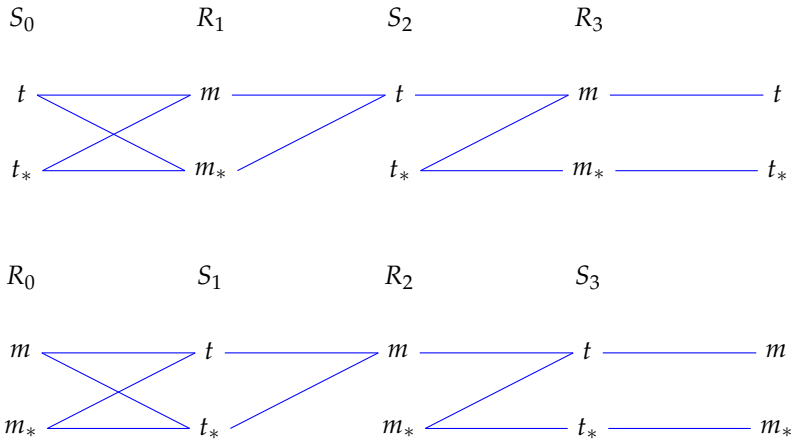


Figure 15: Predictions for Horn's "Division of Pragmatic Labor"

solution concepts have in singling out the proper strategy profile in this kind of game (cf. van Rooij 2004; de Jaegher 2008).

9.3 Some Problems & Some Extensions

So far, the IBR model has done a fair job in explaining the data in accordance with intuition, all of this derived from "first principles" in a formally rigorous account of rational language use and interpretation. Not bad. Still, the IBR model is not flawless, of course. This section reviews some of the problems of this approach, together with some thoughts on how these could be overcome.

Scaling Up. Up to now we have only dealt with disjunctive constructions with two disjuncts. But what happens if we have more than two? Do we still derive correct predictions when we look at plain disjunctions of the form $A_1 \vee \cdots \vee A_n$, FC-disjunctions of the form $\Diamond(A_1 \vee \cdots \vee A_n)$ and SDA-disjunctions of the form $(A_1 \vee \cdots \vee A_n) > C$ when $n > 2$? The answer is a hesitant "yes-

and-no". It is a "yes-and-no" because some predictions scale up smoothly, others don't. It is hesitant because I actually do not believe that we necessarily have to ask this question in the first place (or, rather: ask it and hope for an affirmative answer). Let me enlarge briefly, on the "yes-and-no", sketching some of the results, and then comment on why I don't believe we need to worry too much about long lists of disjuncts in IBR.

Some correct predictions indeed scale up smoothly, for instance, the ignorance and exclusivity implicatures of plain disjunctions. Other intuitive results do not scale up easily. Without additional assumptions in the context model, the IBR model predicts that $\Diamond(A \vee B \vee C)$, for instance, is either a surprise message (in the sequence starting with R_0) or interpreted as "exactly two options out of $\{A, B, C\}$ are allowed" (in the sequence after S_0). These predictions are clearly nonsensical. At the heart of the problem for the kind of "distributional pragmatics" applied by IBR is that the truth-conditional semantics of disjunctions are really rather weak — just as we have noted above in the previous section. With two plausible extra assumptions, however, the IBR model predicts the correct reading for $\Diamond(A_1 \vee \cdots \vee A_n)$. For that we only need to assume that (i) disjunctions are costly proportional to the number of disjuncts and that (ii) states are more likely the fewer alternatives they make true (e.g., the fewer alternatives are permitted). Similar assumptions are also needed for scaling-up the derivation of SDA. The main problem for this case is that already for $n = 3$ we need to distinguish 25 base-level states, some of which impede the intuitive associations of states and message. (Some of these distinctions are clearly spurious and it would be welcome to have a principled motivation for ruling some of these out — see below.)

Be that as it may, I don't think we should worry about this too much. As we have just seen, context models and pragmatic reasoning can get *enormously complicated* as n grows bigger.[43] However, it is not necessary —even from a very conservative Gricean point of view— to assume that a rationalistic explanation of a general inference pattern cannot also make recourse to the idea that language users reason in detail about simple cases and *extrapolate* to complex cases. This extrapolation, this generalization from simple to complex, is not an explicit part of the IBR model, but it would make a feasible Gricean companion to the IBR model. Seen from this perspective, we could say that IBR does a fair part of the ground work, but needs to be backed up with an account of pattern recognition and carry-over inferences of a general kind.

Spurious State Distinctions. But just to delegate models with too many state distinctions to other models of competence (linguistic or otherwise) does

[43]Notice that this as such is not a problem for the IBR model alone. It does not get easier for exhaustification-based approaches either to calculate predictions for larger sets of alternatives.

not suffice. Spurious state distinctions may also come in the way of other, simpler inferences that the IBR model really *should* be able to deal with. For example, Chemla (2009) discusses the case in (48) which is problematic for standard Gricean approaches (cf. Geurts and Pouscoulous 2009*b*; van Rooij 2010).

(48) Everybody is allowed to take an apple or a pear. $\forall x \Diamond (Ax \vee Bx)$

(49) a. Everybody is allowed to take a pear. $\forall x \Diamond Ax$

b. Everybody is allowed to take an apple. $\forall x \Diamond Bx$

It seems fairly intuitive that, at least for deontic modality, the implicatures in (49) should be derivable from (48).[44] But can the IBR model do it?

The base-level interpretation game for this case would distinguish four states:

	$m_{\forall x \Diamond Ax}$	$m_{\forall x \Diamond Bx}$	$m_{\forall x \Diamond (Ax \vee Bx)}$
$t_{[1,0,1]}$	$\surd$	$-$	$\surd$
$t_{[0,1,1]}$	$-$	$\surd$	$\surd$
$t_{[1,1,1]}$	$\surd$	$\surd$	$\surd$
$t_{[0,0,1]}$	$-$	$-$	$\surd$

With this the IBR reasoning is as exemplified in Figure 16: we derive that (48) is associated with the state $t_{[0,0,1]}$, which means that we derive the implicature that the group is mixed: some folks may take an apple but no pear, some others may take a pear but no apple. This is not the intuitively correct prediction.

However, in this particular case, a reasonable means of pruning the context model is ready-at-hand. Notice that the state $t_{[0,0,1]}$ *itself* may be fairly unreasonable: it may be deemed very unlikely in context, or even completely ruled out as an alive possibility. This is not necessarily a dirty trick or a technical hack: it may be a standing presupposition in a natural context of utterance which we construct for sentences like (48) that everybody has *equal rights*. We could implement this assumption in either of two ways: either we could assign to state $t_{[0,0,1]}$ an extremely low, but positive probability, or we could scrap it entirely as an alive option from the context representation. The effect is the same, but the latter affords less computation to spell out the example. Without $t_{[0,0,1]}$ we simply get:

[44] Geurts and Pouscoulous (2009*b*) argue that implicatures of this kind are not as natural for existential operators other than deontic modals. This observation fits in nicely with the account given here.

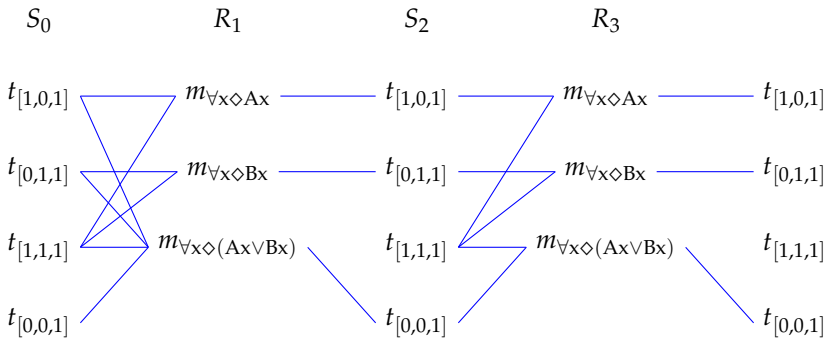
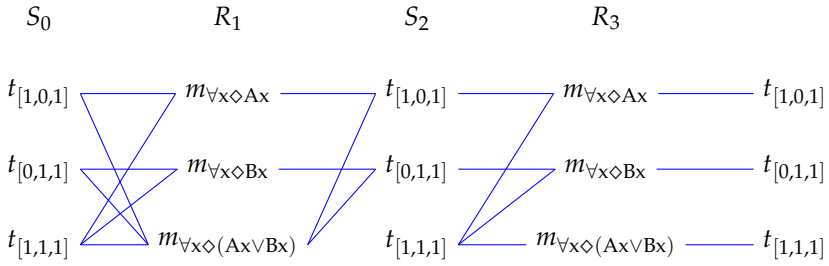


Figure 16: Predictions for example (48) in full context model



This is the desired implicature, which we derive —in Gricean fashion— whenever an assumption of "group homogeneity" is feasible simply by pruning the set of state distinctions. This demonstrates that it is reasonable and beneficial to occasionally prune or alter the context model in certain ways. But saying this much is still far removed from the necessary principled and general specification of exactly which alterations and extra assumptions should be added under exactly which circumstances — an open end for further scrutiny.

10 Conclusion

Taking stock, this paper has offered a general game-theoretic model of quantity implicature calculation. The model consists of two parts: (i) a general procedure with which to construct interpretation games as models of the context of utterance from a set of alternative sentences, and (ii) a step-by-step reasoning process that selects the pragmatically feasible play in these games. This approach deals uniformly with a variety of quantity implicatures, and is also versatile enough to make general and yet flexible predictions about different strength of epistemic quantity implicatures.

The model's predictions are given by fixed points of ever more sophisticated theory-of-mind reasoning. Central results established that fixed points are always reached for interpretation games, and that these correspond to perfect Bayesian equilibria. The paper also showed that we can compute IBR reasoning by a manageable algorithm, if parameters (such as message costs and prior probabilities) are adequately chosen. Of course, generally a game-theoretic approach to pragmatic reasoning is much more powerful than the specialized model of this paper. The present model would straightforwardly extend also to cases, for instance, where the speaker's and the hearer's interests are in conflict, or where message costs are severe.

The model given in this paper is superficially very similar to (iterated) exhaustive interpretation. It subsumes exhaustification in terms of minimal models as a special case in all examples that we looked at in this paper, namely as rational interpretation of a hearer who merely takes the distributional information given by the semantic meanings of a set of alternative expressions into account. The model, however, deviates from the predictions of iterated applications of Exh_{MM} in that it is not monotone: unlike the latter, the former reconsiders all options (formulations or interpretations) anew at each iteration step. This way the "pragmatically proper" mappings between formulations and interpretations may be established at later iterations, even if excluded at earlier steps.

But whereas the IBR model bears a clear relationship to iterated exhaustification based on minimal models, there is also an uncomfortable gap in the picture sketched in this paper. The IBR model does not relate in the same obvious way to exhaustive interpretation in terms of innocent exclusion. It remains unclear from the present perspective if and how this interpretation operation could be explained as a rational inference or, more generally, as the outcome of some process optimizing speaker's and hearer's interests in successful communication. Since Exh_{IE} predicts formidably for many cases, but not for SDA, one could hope that by probing for its rationale, this interpretation operator could be improved to also cover SDA. Presently, however, I do not see how, and must leave this for future consideration.

A Comparison of Exhaustivity Operators

This section compares the semantic (minimal-models) approach to exhaustive interpretation with the syntactic (innocent-exclusion) approach (see Section 4 for definitions). Three results characterize the general relationship between these approaches: (i) Fact A.1 states that Exh_{MM} is insensitive to certain variations in the set Alt, (ii) Fact A.2 establishes that the interpretation selected by Exh_{MM} always entails that selected by Exh_{IE}, and (iii) Fact A.4 gives sufficient and necessary conditions on Alt for identity of predictions.

To start with, here is a simple example that shows that the operators Exh_{MM} and Exh_{IE} are not generally equivalent. Consider a disjunction with two (logically independent) disjuncts and two candidate sets of alternatives.

(50) A or B.

(51) a. $\text{Alt}_1 = \{A,\ B,\ A \vee B\}$

b. $\text{Alt}_2 = \text{Alt}_1 \cup \{A \wedge B\}$

Let's calculate the predictions for both approaches, starting with the approach in terms of minimal models. We only need to consider three types of worlds that are true in S: w_A where A is true and B is false, w_B where B is true and A is false and w_{AB} where both A and B are true. The minimal worlds in this triple with respect to both $<_{\text{Alt}_1}$ and $<_{\text{Alt}_2}$ are all worlds of type w_A or of type w_B. In other words, the additional conjunctive alternative in Alt_2 does not influence the ordering on possible worlds. Hence, the predictions of the minimal-models approach are the same for both sets of alternatives:

$$\text{Exh}_{\text{MM}}(A \vee B, \text{Alt}_1) = \text{Exh}_{\text{MM}}(A \vee B, \text{Alt}_2) = \{w_A, w_B\}\,.$$

This is different for the innocent-exclusion approach. Consider first Alt_1. The maximal consistently excludable sets are $\{A\}$ and $\{B\}$, but their intersection is empty, so that

$$\text{Exh}_{\text{IE}}(A \vee B, \text{Alt}_1) = A \vee B = \{w_A, w_B, w_{AB}\}\ .$$

Considering instead Alt_2, the maximal consistently excludable sets are $\{A,\ A \wedge B\}$ and $\{B,\ A \wedge B\}$. These have a non-empty intersection so that

$$\text{Exh}_{\text{IE}}(A \vee B, \text{Alt}_2) = \{w_A, w_B\}\ .$$

This is equivalent to the prediction of the semantic approach.

Taken together, the presence or absence of the conjunctive alternative matters to the syntactic, but not to the semantic approach. The latter predicts *as if* the conjunctive alternative was given. Schematically:

$$\text{Exh}_{\text{MM}}(\text{Alt}_1) = \text{Exh}_{\text{MM}}(\text{Alt}_2) = \text{Exh}_{\text{IE}}(\text{Alt}_2) \subset \text{Exh}_{\text{IE}}(\text{Alt}_1)\,.$$

We can generalize the point of this example. Indeed, the predictions of the semantic approach are necessarily equivalent under a certain variance in the set ALT; the syntactic approach does not have this invariance property. Notice that in the example above, the orderings $<_{\text{ALT}_1}$ and $<_{\text{ALT}_2}$ were identical because each world that assigns truth values to propositions A and B *thereby* also assigns a truth value to $A \wedge B$. Let's say that a proposition A is *truth-determined* by a set of propositions X if the truth value of A is completely determined by any truth value assignment to all members of X. The ordering $<_X$ is invariant under addition or removal of truth-determined alternatives in the following sense:

Fact A.1. If A is truth-determined by X, then $<_X = <_{X \cup \{A\}}$.

This means that EXH_{MM} is less sensitive to the exact specification of the alternatives: if certain propositions A and B are alternatives in ALT, then EXH_{MM} also, as it were, implicitly considers the conjunctive alternative $A \wedge B$. In contrast, EXH_{IE} does not.

This suggests that pragmatic interpretation in terms of EXH_{MM} is always included —but not necessarily strictly— in the pragmatic interpretation in terms of EXH_{IE}. This is borne out in general:

Fact A.2. For any S and ALT, we have $\text{EXH}_{\text{MM}}(S, \text{ALT}) \subseteq \text{EXH}_{\text{IE}}(S, \text{ALT})$.

Proof. Take an arbitrary $w \in \text{EXH}_{\text{MM}}(S, \text{ALT})$. This means that w is minimal in S with respect to ordering $<_{\text{ALT}}$, which in turn means that w makes a maximal number of alternatives in ALT false. In other words, there is an $\mathcal{A} \in \text{Max-CE}(S, \text{ALT})$ such that $w \in S \wedge \bigwedge_{A \in \mathcal{A}} \neg A$. But then w is also in any proposition $S \wedge \bigwedge_{A \in \mathcal{B}} \neg A$ for $\mathcal{B} \subseteq \mathcal{A}$. In particular, then, $w \in \text{EXH}_{\text{IE}}(S, \text{ALT})$. □

Next, we would like to know under which circumstances exactly the operators coincide. This, obviously, hinges on the set ALT. While EXH_{MM} rules out all non-minimal worlds according to $<_{\text{ALT}}$, the operator EXH_{IE} does not necessarily rule out all non-minimal worlds, but only those worlds $w \in S$ for which there is an alternative $A \in \text{ALT}$ such that *all* minimal worlds make A true (resp. false), while w makes A false (resp. true). In other terms, EXH_{IE} is more conservative than EXH_{MM} in that it selects as pragmatic interpretation of S all those worlds in S which cannot be distinguished from the minimal worlds by some alternative in ALT. To formalize this, we introduce a suitable notion of distinguishability: we say that world w is ALT-distinguishable from the set of worlds $X \subseteq W$ iff there is some $A \in \text{ALT}$ such that either all worlds in X make A true while w makes A false, or all worlds in X make A false while

w makes A true. If w is ALT-indistinguishable from set X, write $w \sim_{\text{ALT}} X$. These considerations give rise to the following:

Lemma A.3. $\text{EXH}_{\text{IE}}(S, \text{ALT}) = \{w \in S \mid w \sim_{\text{ALT}} \text{EXH}_{\text{MM}}(S, \text{ALT})\}$

In other terms, we can characterize EXH_{IE} semantically as the closure of EXH_{MM} under ALT-indistinguishability.

So when exactly does a set ALT have the property that the minimal worlds according to $<_{\text{ALT}}$ are closed under ALT-indistinguishability? To state sufficient and necessary conditions for this formally, define for all $w \in \text{EXH}_{\text{MM}}$ that $\text{A}(w)$ is the unique strongest proposition in $\text{ALT} \cup \{S\}$ that is true in w. With this, we can state the following sufficient and necessary condition on ALT for equivalence of EXH_{MM} and EXH_{IE}.

Fact A.4. $\text{EXH}_{\text{MM}}(S, \text{ALT}) = \text{EXH}_{\text{IE}}(S, \text{ALT})$ iff for all $w, w' \in \text{EXH}_{\text{MM}}(S, \text{ALT})$ there is an alternative $A \in \text{ALT}$ such that the conjunction of $\text{A}(w)$ and $\text{A}(w')$ entails A.

Proof. With Lemma A.3 it suffices to show that the right hand side of Fact A.4 is equivalent to:

$$\text{EXH}_{\text{MM}}(S, \text{ALT}) = \{w \in S \mid w \sim_{\text{ALT}} \text{EXH}_{\text{MM}}(S, \text{ALT})\} \ .$$

This is so because, firstly, if for all worlds w, w' in $\text{EXH}_{\text{MM}}(S, \text{ALT})$ there is an A with the designated property, then we *can* distinguish all non-minimal worlds from the minimal ones. Secondly, if for a given pair of worlds w, w' in $\text{EXH}_{\text{MM}}(S, \text{ALT})$ there is no such A with the designated property, then there is a world w^* which makes both $A(w)$ and $A(w')$ true and all other $A \in \text{ALT}$ false. This w^* is not minimal, but it is also not ALT-distinguishable from the set $\{w, w'\}$ and therefore also not from $\text{EXH}_{\text{MM}}(S, \text{ALT})$. □

B Proofs

B.1 Proof of Theorem 8.1

Definition B.1 (Pure Cooperation, Expected Gain). A *signaling game of pure cooperation* is one where sender and receiver utilities are aligned: $\text{U}_S(t, m, a) = \text{U}_R(t, m, a)$ for all t, m, a. For these games, we define the *expected gain* of a pair of strategies σ, ρ as:

$$\text{EG}(\sigma, \rho) = \sum_t \Pr(t) \times \sum_m \sigma(t, m) \times \sum_a \rho(m, a) \times \text{U}(t, m, a) \, .$$

Lemma B.2. In a signaling game of pure cooperation the expected gain is monotone increasing along the IBR sequence, in the sense that for all $i \geq 0$:

(i) $\mathrm{EG}(S_i, R_{i+1}) \leq \mathrm{EG}(S_{i+2}, R_{i+1})$, and

(ii) $\mathrm{EG}(S_{i+1}, R_i) \leq \mathrm{EG}(S_{i+1}, R_{i+2})$.

Proof. Ad (i). It holds for all t that:

$$S_{i+2}(t) \subseteq \arg\max_{m \in M} \sum_a R_{i+1}(m,a) \times \mathrm{U}_S(t,m,a)\,.$$

This implies that for all $t \in T$:

$$\begin{aligned} &\sum_m S_i(t,m) \times R_{i+1}(m,a) \times \mathrm{U}(t,m,a) \\ \leq &\sum_m S_{i+2}(t,m) \times R_{i+1}(m,a) \times \mathrm{U}(t,m,a)\,. \end{aligned}$$

This, in turn, implies:

$$\begin{aligned} &\sum_t \Pr(t) \sum_m S_i(t,m) \times R_{i+1}(m,a) \times \mathrm{U}(t,m,a) \\ \leq &\sum_t \Pr(t) \sum_m S_{i+2}(t,m) \times R_{i+1}(m,a) \times \mathrm{U}(t,m,a)\,. \end{aligned}$$

And this is equivalent to $\mathrm{EG}(S_i, R_{i+1}) \leq \mathrm{EG}(S_{i+2}, R_{i+1})$.

Ad (ii). Begin by rewriting the statement to be shown:

$$\begin{aligned} & \mathrm{EG}(S_{i+1}, R_i) \leq \mathrm{EG}(S_{i+1}, R_{i+2}) \\ \text{iff} \quad & \sum_t \Pr(t) \times \sum_m S_{i+1}(t,m) \times \sum_a R_i(m,a) \times \mathrm{U}_{S,R}(t,m,a) \leq \\ & \sum_t \Pr(t) \times \sum_m S_{i+1}(t,m) \times \sum_a R_{i+2}(m,a) \times \mathrm{U}_{S,R}(t,m,a) \\ \text{iff} \quad & \sum_t \sum_m \sum_a \Pr(t) \times S_{i+1}(t,m) \times R_i(m,a) \times \mathrm{U}_{S,R}(t,m,a) \leq \\ & \sum_t \sum_m \sum_a \Pr(t) \times S_{i+1}(t,m) \times R_{i+2}(m,a) \times \mathrm{U}_{S,R}(t,m,a)\,. \end{aligned}$$

Observe that for messages that surprise R_{i+2} we have $S_{i+1}(t,m) = 0$ for all t, so that the receiver's reception of these messages does not figure in the inequality. Let $M^* = \left\{ m \in M \mid S_{i+1}^{-1}(m) \neq \emptyset \right\}$ be the set of non-surprise messages under S_{i+1}. The previous statement therefore is equivalent to:

$$\begin{aligned} &\sum_t \sum_{m \in M^*} \sum_a \Pr(t) \times S_{i+1}(t,m) \times R_i(m,a) \times \mathrm{U}_{S,R}(t,m,a) \leq \\ &\sum_t \sum_{m \in M^*} \sum_a \Pr(t) \times S_{i+1}(t,m) \times R_{i+2}(m,a) \times \mathrm{U}_{S,R}(t,m,a)\,. \end{aligned}$$

Dividing each summand on both sides with a constant $0 \neq c(m) = \sum_{t'} \Pr(t') \times S_{i+1}(t', m)$ for each $m \in M^*$ yields:

$$\sum_t \sum_{m \in M^*} \sum_a \frac{\Pr(t) \times S_{i+1}(t,m)}{c(m)} \times R_i(m,a) \times \mathrm{U}_{S,R}(t,m,a) \leq$$
$$\sum_t \sum_{m \in M^*} \sum_a \frac{\Pr(t) \times S_{i+1}(t,m)}{c(m)} \times R_{i+2}(m,a) \times \mathrm{U}_{S,R}(t,m,a)$$

$$\text{iff} \quad \sum_t \sum_{m \in M^*} \sum_a \mu_{i+1}(t|m) \times R_i(m,a) \times \mathrm{U}_{S,R}(t,m,a) \leq$$
$$\sum_t \sum_{m \in M^*} \sum_a \mu_{i+1}(t|m) \times R_{i+2}(m,a) \times \mathrm{U}_{S,R}(t,m,a)\,.$$

We now see that this inequality holds, because for any $m \in M^*$ we have $R_{i+2}(m) \subseteq \arg\max_a \mu_{i+2}(t|m) \times \mathrm{U}_S(t,m,a)\,.$ □

Proof of Theorem 8.1. By Lemma B.2 we know that expected gain is monotone increasing. Clearly, $\mathrm{EG}(\cdot)$ is upper-bounded for finite games. Since there are also only finitely many sets of pure strategies (that could constitute types in an IBR sequence), and since the IBR sequence is entirely deterministic, each sequence must reach a highest value for $\mathrm{EG}(\cdot)$. This entails a fixed point, because from $\mathrm{EG}(S_i, R_{i+1}) = \mathrm{EG}(S_{i+2}, R_{i+1})$, it follows that

$$S_{i+2}(t) \subseteq \arg\max_{m \in M} \sum_a R_{i+1}(m,a) \times \mathrm{U}_S(t,m,a)\,,$$

and this implies that $S_i(t) \subseteq S_{i+2}(t)$ for all t. But for finite set M, this cannot be an infinite sequence with a strict subset relation. □

B.2 Proof of Theorem 8.2

Definition B.3 (Perfect Bayesian Equilibrium). We say that a triple $\langle \sigma, \rho, \mu \rangle \in (\Delta(M))^T \times (\Delta(A))^M \times (\Delta(T))^M$ is a PERFECT BAYESIAN EQUILIBRIUM (PBE) iff three conditions hold:[45]

(i) σ is rational given the belief ρ;

(ii) ρ is rational given the belief μ;

(iii) μ is consistent with Pr and the belief σ.

We say that a strategy profile $\langle \sigma, \rho \rangle$ *gives rise to* a PBE iff there is a posterior μ such that $\langle \sigma, \rho, \mu \rangle$ is a PBE.

[45]Strictly speaking, the notion of rationality of a *probabilistic* strategy under an *arbitrary* behavioral belief has not been defined in this paper, but it bears no surprises. The interested reader is referred to standard textbooks.

Proposition B.4. If $\langle S^*, R^* \rangle$ is a fixed point of an IBR sequence such that there are no surprise messages under S^*, then $\langle S^*, R^* \rangle$ gives rise to a perfect Bayesian equilibrium.

Proof. If $\langle S^*, R^* \rangle$ is the fixed point of an IBR sequence, S^* is a best response to the belief R^*. Moreover, if there are no surprise messages under S^*, then there is only one posterior belief μ^* consistent with the given prior and the belief S^*. By definition of IBR types, R^* is a best response to μ^*. Hence, all conditions for perfect Bayesian equilibrium are fulfilled by the triple $\langle S^*, R^*, \mu^* \rangle$. □

Proof of Theorem 8.2. Given Proposition B.4 and its proof, we only need to show that for any fixed point $\langle S^*, R^* \rangle$ there is a posterior μ under which $R^*(m)$ is rational for surprise messages m. Given conditions C1 and C2 this is fulfilled by the unique μ^* which is consistent with S^*, and for which for all surprise messages m we have: $\mu^*(t|m) = |[\![m]\!]|^{-1}$. □

B.3 Proof of Theorem 8.3

Lemma B.5. Under the conditions C1 and C2 from Theorems 8.2 and 8.3 truth is preserved by all types: $S_k(t) \subseteq [\![t]\!]^{-1}$ and $R_k(m) \subseteq [\![m]\!]$ for all k.

Proof of Lemma B.5. By induction. The base case is trivial. So suppose that $S_k(t) \subseteq [\![t]\!]^{-1}$ for all t, and show $R_{k+1}(m) \subseteq [\![m]\!]$, for all m. First, take the case $S_k^{-1}(m) \neq \emptyset$, i.e., a non-surprise message m. By inductive hypothesis $S_k^{-1}(m) \subseteq [\![m]\!]$, and so together with conditions C1 and C2:

$$R_{k+1}(m) = \arg\max_{t \in T} \mu_{k+1}(t|m) \subseteq S_k^{-1}(m) \subseteq [\![m]\!] \ .$$

In case of a surprise message with $S_k^{-1}(m) = \emptyset$, any state is a best response, and so by TCP assumption $R_{k+1}(m) \subseteq [\![m]\!]$.

The induction step for the sender is almost identical. Suppose that $R_k(m) \subseteq [\![m]\!]$ for all m, and show $S_{k+1}(t) \subseteq [\![t]\!]^{-1}$ for all t. First, take the case $R_k^{-1}(t) \neq \emptyset$. By induction hypothesis, C1 and C2:

$$S_{k+1}(t) = \arg\max_{m \in M} \sum_{a \in A} R_k(m, a) \times \mathrm{U}_S(t, m, a) \subseteq R_k^{-1}(t) \subseteq [\![t]\!]^{-1} \ .$$

In case $R_k^{-1}(t) = \emptyset$, any message maximizes expected utility given R_k, so that by TCP assumption $S_{k+1}(t) \subseteq [\![t]\!]^{-1}$. □

Proof of Theorem 8.3. By induction. As the base case is trivial, assume first that $R_k = \check{R}_k$ and show that $S_{k+1} = \check{S}_{k+1}$. By definition:

$$S_{k+1} = \left\{ s \in \mathrm{BR}(R_k) \mid \forall t \ (\exists s' \in \mathrm{BR}(R_k) \ t \in [\![s'(t)]\!]) \rightarrow t \in [\![s(t)]\!] \right\} \ .$$

First, take a state for which $R_k^{-1}(t) \neq \emptyset$. From Lemma B.5, we know that $R_k(m) \subseteq [\![m]\!]$, so that:

$$S_{k+1}(t) = \arg\max_{m \in M} \sum_{a \in A} R_k(m,a) \times \mathrm{U}_S(t,m,a)\,.$$

This expected utility boils down to the following under assumptions C1 and C2:

$$\begin{aligned}\sum_{a \in A} R_k(m,a) \times \mathrm{U}_S(t,m,a) &= \sum_{t' \in T} R_k(m,t') \times \mathrm{U}_S(t,m,t') \\ &= \begin{cases} \frac{1}{|R_k(m)|} & \text{if } t \in R_k(m) \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

From the last two equations, $S_{k+1}(t) = \breve{S}_{k+1}(t)$ follows. If, on the other hand, $R_k^{-1}(t) = \emptyset$, then any message will maximize expected utility, and the speaker will, by TCP assumption, send arbitrarily any true message, so that $S_{k+1}(t) = [\![t]\!]^{-1} = \breve{S}_{k+1}(t)$.

Finally, assume that $S_k = \breve{S}_k$ and show that $R_{k+1} = \breve{R}_{k+1}$. For surprise messages with $S_k(m)^{-1} = \emptyset$, any state maximizes expected utility given some consistent belief in S_k. In that case, by TCP assumption $R_{k+1}(m) = \breve{R}_{k+1}(m)$. So, suppose m is not a surprise, i.e., $S_k(m)^{-1} \neq \emptyset$. Then Bayesian conditionalization gives a unique $\mu_{k+1}(\cdot|m)$ for which by C1 and C2:

$$R_k(m) = \arg\max_{t \in T} \mu_{k+1}(t|m)\,.$$

To solve this maximization, calculate:

$$\begin{aligned} & & \mu_{k+1}(t_1|m) > \mu_{k+1}(t_2|m) \\ & \text{iff} & \frac{\Pr(t_1) \times S_k(t_1,m)}{\sum_{t' \in T} \Pr(t') \times S_k(t',m)} > \frac{\Pr(t_2) \times S_k(t_2,m)}{\sum_{t' \in T} \Pr(t') \times S_k(t',m)} \\ & \text{iff} & \Pr(t_1) \times S_k(t_1,m) > \Pr(t_2) \times S_k(t_2,m) \\ \text{(from C3)} & \text{iff} & S_k(t_1,m) > S_k(t_2,m) \\ & \text{iff} & \big(m \in S_k(t_1) \text{ and } m \notin S_k(t_2)\big) \text{ or} \\ & & \big(m \in S_k(t_1) \text{ and } m \in S_k(t_2) \text{ and } |S_k(t_1)| < |S_k(t_2)|\big) \end{aligned}$$

With this, it is clear that $R_k(m) = \breve{R}_k(m)$. □

References

Allott, Nicholas (2006). "Game Theory and Communication". In: *Game Theory and Pragmatics*. Ed. by Anton Benz et al. Palgrave MacMillan. Pp. 123–151.

Alonso-Ovalle, Luis (2005). "Distributing the Disjuncts over the Modal Space". In: *Proceedings of the North East Linguistics Society*. Ed. by Leah Bateman and Cherlon Ussery. Vol. 35. GLSA. Amherst, MA.

— (2008). "Innocent Exclusion in an Alternative-Semantics". In: *Natural Language Semantics* 16.2. Pp. 115–128.

Asher, Nicholas and Daniel Bonevac (2005). "Free Choice Permission as Strong Permission". In: *Synthese* 145.3. Pp. 303–323.

Atlas, Jay David and Stephen Levinson (1981). "It-clefts, Informativeness, and Logical Form". In: *Radical Pragmatics*. Ed. by Peter Cole. Academic Press. Pp. 1–61.

Bach, Kent (2006). "The Top 10 Misconceptions about Implicature". In: *Drawing the Boundaries of Meaning: Neo-Gricean Studies in Pragmatics and Semantics in Honor of Laurence R. Horn*. Ed. by Betty Birner and Gregory Ward. John Benjamins. Pp. 21–30.

Barker, Chris (2010). "Free Choice Permission as Resource-Sensitive Reasoning". In: *Semantics & Pragmatics* 3.10. Pp. 1–38.

Benz, Anton and Robert van Rooij (2007). "Optimal Assertions and what they Implicate". In: *Topoi* 26. Pp. 63–78.

Benz, Anton et al., eds. (2006). *Game Theory and Pragmatics*. Hampshire: Palgrave MacMillan.

Block, Eliza (2008). "Is the Symmetry Problem Really a Problem?" Unpublished manuscript, NYU.

Camerer, Colin F. (2003). *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press.

Camerer, Colin F. et al. (2004). "A Cognitive Hierarchy Model of Games". In: *The Quarterly Journal of Economics* 119.3. Pp. 861–898.

Chapman, Siobhan (2005). *Paul Grice, Philosopher and Linguist*. Hampshire: Palgrave MacMillan.

Chemla, Emmanuel (2009). "Universal Implicatures and Free Choice Effects: Experimental Data". In: *Semantics & Pragmatics* 2.2. Pp. 1–33.

Chierchia, Gennaro (2004). "Scalar Implicatures, Polarity Phenomena and the Syntax/Pragmatics Interface". In: *Structures and Beyond*. Ed. by Adriana Belletti. Oxford University Press. Pp. 39–103.

Chierchia, Gennaro et al. (2008). "The Grammatical View of Scalar Implicatures and the Relationship between Semantics and Pragmatics". Unpublished manuscript.

Cho, In-Koo and David M Kreps (1987). "Signaling Games and Stable Equilibria". In: *The Quarterly Journal of Economics* 102.2. Pp. 179–221.

Crawford, Vincent P. (2007). "Let's Talk It Over: Coordination Via Preplay Communication With Level-k Thinking". Unpublished manuscript.

Crawford, Vincent P. and Nagore Iriberri (2007). "Fatal Attraction: Salience, Naïveté, and Sophistication in Experimental "Hide-and-Seek" Games". In: *The American Economic Review* 97.5. Pp. 1731–1750.

Culicover, Peter W. and Ray Jackendoff (1997). "Semantic subordination despite syntactic coordination". In: *Linguistic Inquiry* 28.2. Pp. 195–217.

Farrell, Joseph (1993). "Meaning and Credibility in Cheap-Talk Games". In: *Games and Economic Behavior* 5. Pp. 514–531.

Farrell, Joseph and Matthew Rabin (1996). "Cheap Talk". In: *The Journal of Economic Perspectives* 10.3. Pp. 103–118.

Fine, Kit (1975). "Critical Notice: Counterfactuals." In: *Mind* 84.335. Pp. 451–458.

Fox, Danny (2007). "Free Choice and the Theory of Scalar Implicatures". In: *Presupposition and Implicature in Compositional Semantics*. Ed. by Uli Sauerland and Penka Stateva. Hampshire: Palgrave MacMillan. Pp. 71–120.

Franke, Michael (2008). "Pseudo-Imperatives and Other Cases of Conditional Conjunction and Conjunctive Disjunction". In: *'Subordination' versus 'Coordination' in Sentence and Text — From a Cross-Linguistic Perspective*. Ed. by Cathrine Fabricius-Hansen and Wiebke Ramm. Studies in Language Companion Series (SLCS). Amsterdam, Philadelphia: John Benjamins. Pp. 255–279.

— (2009). "Signal to Act: Game Theory in Pragmatics". PhD thesis. Universiteit van Amsterdam.

— (2010). "Semantic Meaning and Pragmatic Inference in Non-cooperative Conversation". In: *Interfaces: Explorations in Logic, Language and Computation*. Ed. by Thomas Icard and Reinhard Muskens. Lecture Notes in Artificial Intelligence. Berlin, Heidelberg: Springer-Verlag. Pp. 13–24.

Gazdar, Gerald (1979). *Pragmatics: Implicature, Presupposition, and Logical Form*. New York: Academic Press.

Geurts, Bart (2005). "Entertaining Alternatives: Disjunctions as Modals". In: *Natural Language Semantics* 13. Pp. 383–410.

— (2010). *Quantity Implicatures*. Cambridge University Press.

Geurts, Bart and Nausicaa Pouscoulous (2009*a*). "Embedded Implicatures?!?" In: *Semantics & Pragmatics* 2.4. Pp. 1–34.

— (2009*b*). "Free Choice for All: A Response to Emmanuell Chemla". In: *Semantics & Pragmatics* 2.5. Pp. 1–10.

Gómez-Txurruka, Isabell (2002). *The Semantics of Natural Language Disjunction Or*. Unpublished manuscript. ILCLI Donostia-San Sebastián.

Grafen, Alan (1990). "Biological Signals as Handicaps". In: *Journal of Theoretical Biology* 144. Pp. 517–546.

Grice, Paul Herbert (1975). "Logic and Conversation". In: *Syntax and Semantics, Vol. 3, Speech Acts*. Ed. by Peter Cole and Jerry L. Morgan. Academic Press. Pp. 41–58.

— (1989). *Studies in the Ways of Words*. Harvard University Press.

Groenendijk, Jeroen and Martin Stokhof (1984). "Studies in the Semantics of Questions and the Pragmatics of Answers". PhD thesis. Universiteit van Amsterdam.

Ho, Teck-Hua et al. (1998). "Iterated Dominance and Iterated Best Response in Experimental 'p-Beauty Contests' ". In: *The American Economic Review* 88.4. Pp. 947–969.

Horn, Laurence R. (1972). "On the Semantic Properties of Logical Operators in English". PhD thesis. UCLA.

— (1984). "Towards a New Taxonomy for Pragmatic Inference: Q-based and I-based implicatures". In: *Meaning, Form, and Use in Context*. Ed. by Deborah Shiffrin. Washington: Georgetown University Press. Pp. 11–42.

— (1989). *A Natural History of Negation*. Chicago: Chicago University Press.

de Jaegher, Kris (2008). "The Evolution of Horn's Rule". In: *Journal of Economic Methodology* 15.3. Pp. 275–284.

Jäger, Gerhard (2007). "The Evolution of Convex Categories". In: *Linguistics and Philosophy* 30.5. Pp. 551–564.

Jäger, Gerhard and Christian Ebert (2009). "Pragmatic Rationalizability". In: *Proceedings of Sinn und Bedeutung 13*. Ed. by Arndt Riester and Torgrim Solstad. Pp. 1–15.

de Jager, Tikitu and Robert van Rooij (2007). "Explaining Quantity Implicatures". In: *Proceedings of the 11th conference on Theoretical Aspects of Rationality and Knowledge*. New York: ACM. Pp. 193–202.

Kamp, Hans (1973). "Free Choice Permission". In: *Proceedings of the Aristotelian Society* 74. Pp. 57–74.

— (1978). "Semantics versus Pragmatics". In: *Formal Semantics and Pragmatics for Natural Languages*. Ed. by Franz Guenthner and Siegfried Josef Schmidt. Dordrecht: Reidel. Pp. 255–287.

Katzir, Roni (2007). "Structurally-Defined Alternatives". In: *Linguistics and Philosophy* 30.6. Pp. 669–690.

Klinedinst, Nathan (2006). "Plurality and Possibility". PhD thesis. University of California, Los Angeles.

Kratzer, Angelika (1981). "The Notional Category of Modality". In: *Words, Worlds, and Contexts: New Approaches in Word Semantics*. Ed. by H. J. Eikmeyer and H. Rieser. Berlin: de Gruyter. Pp. 38–74.

Kratzer, Angelika and Junko Shimoyama (2002). "Indeterminate Pronouns: The View from Japanese". In: *Proceeding of the 3rd Tokyo Conference on Psycholinguistics*. Ed. by Yukio Otsu. Pp. 1–25.

van Kuppevelt, Jan (1996). "Inferring from Topics: Scalar Implicatures as Topic-Dependent Inferences". In: *Linguistics and Philosophy* 19.4. Pp. 393–443.

Levinson, Stephen C. (1983). *Pragmatics*. Cambridge, UK: Cambridge University Press.

— (2000). *Presumptive Meanings. The Theory of Generalized Conversational Implicature*. Cambridge, Massachusetts: MIT Press.

Lewis, David (1969). *Convention. A Philosophical Study*. Harvard University Press.

— (1973). *Counterfactuals*. Harvard University Press.

— (1981). "Ordering Semantics and Premise Semantics for Counterfactuals". In: *Journal of Philosophical Logic* 10.2. Pp. 217–234.

Matsumoto, Yo (1995). "The Conversational Condition on Horn Scales". In: *Linguistics and Philosophy* 18.1. Pp. 21–60.

McClure, William (2000). *Using Japanese — A Guide to Contemporary Usage*. Cambridge University Press.

McKay, Thomas and Peter van Inwagen (1977). "Counterfactuals with Disjunctive Antecedents". In: *Philosophical Studies* 31.5. Pp. 353–356.

Merin, Arthur (1992). "Permission Sentences Stand in the Way of Boolean and Other Lattice-Theoretic Semantices". In: *Journal of Semantics* 9. Pp. 95–162.

Mühlenbernd, Roland (2009). "Kommunikationsmodell für den Entwicklungsprozess von Implikaturen". MA thesis. University of Bielefeld.

Nute, Donald (1975). "Counterfactuals and the Similarity of Worlds". In: *Journal of Philosophy* 21.4. Pp. 773–778.

Parikh, Prashant (2001). *The Use of Language*. Stanford University: CSLI Publications.

Rabin, Matthew (1990). "Communication between Rational Agents". In: *Journal of Economic Theory* 51. Pp. 144–170.

Roberts, Craige (1989). "Modal Subordination and Pronominal Anaphora in Discourse". In: *Linguistics and Philosophy* 12. Pp. 683–721.

van Rooij, Robert (2000). "Permission to Change". In: *Journal of Semantics* 17.2. Pp. 119–143.

— (2003). "Questioning to Resolve Decision Problems". In: *Linguistics and Philosophy* 29. Pp. 727–763.

— (2004). "Signalling Games Select Horn-Strategies". In: *Linguistics and Philosophy* 27. Pp. 493–527.

— (2006). "Free Choice Counterfactual Donkeys". In: *Journal of Semantics* 23.4. Pp. 383–402.

van Rooij, Robert (2008). "Games and Quantity Implicatures". In: *Journal of Economic Methodology* 15.3. Pp. 261–274.
— (2010). "Conjunctive Interpretation of Disjunction". In: *Semantics & Pragmatics* 3.11. Pp. 1–28.
van Rooij, Robert and Katrin Schulz (2004). "Exhaustive Interpretation of Complex Sentences". In: *Journal of Logic, Language and Information* 13. Pp. 491–519.
— (2006). "Only: Meaning and Implicatures". In: *Questions in Dynamic Semantics*. Ed. by Maria Aloni et al. Amsterdam, Singapore: Elsevier. Pp. 193–223.
Ross, Alf (1941). "Imperatives and Logic". In: *Theoria* 25.7. Pp. 53–71.
Russell, Benjamin (2006). "Against Grammatical Computation of Scalar Implicatures". In: *Journal of Semantics* 23.361–382.
Sauerland, Uli (2004). "Scalar Implicatures in Complex Sentences". In: *Linguistics and Philosophy* 27. Pp. 367–391.
Schulz, Katrin (2005). "A Pragmatic Solution for the Paradox of Free Choice Permission". In: *Synthese* 147. Pp. 343–377.
Schulz, Katrin and Robert van Rooij (2006). "Pragmatic Meaning and Nonmonotonic Reasoning: The Case of Exhaustive Interpretation". In: *Linguistics and Philosophy* 29. Pp. 205–250.
Simons, Mandy (2005). "Dividing Things Up: The Semantics of *or* and the modal/*or* interaction". In: *Natural Language Semantics* 13.3. Pp. 271–316.
Soames, Scott (1982). "How Presuppositions are Inherited: A Solution to the Projection Problem". In: *Linguistic Inquiry* 13.3. Pp. 483–545.
Spector, Benjamin (2006). "Scalar Implicatures: Exhaustivity and Gricean Reasoning". In: *Questions in Dynamic Semantics*. Ed. by Maria Aloni et al. Amsterdam, Singapore: Elsevier. Pp. 229–254.
— (2007). "Aspects of the Pragmatics of Plural Morphology: On Higher-Order Implicatures". In: *Presupposition and Implicature in Compositional Semantics*. Ed. by Uli Sauerland and Penka Stateva. Palgrave MacMillan. Pp. 243–281.
Spence, Andrew Michael (1973). "Job market signaling". In: *Quarterly Journal of Economics* 87. Pp. 355–374.
Stalnaker, Robert (1968). "A Theory of Conditionals". In: *Studies in Logical Theory*. Ed. by Nicholas Rescher. Vol. 2. Oxford University Press. Pp. 98–112.
— (1998). "Belief Revision in Games: Forward and Backward Induction". In: *Mathematical Social Sciences* 36. Pp. 31–56.
— (2006). "Saying and Meaning, Cheap Talk and Credibility". In: *Game Theory and Pragmatics*. Ed. by Anton Benz et al. Hampshire: Palgrave MacMillan. Pp. 83–100.
von Stechow, Arnim and Thomas Ede Zimmermann (1984). "Term Answers and Contextual Change". In: *Linguistics* 22.1. Pp. 3–40.

Swanson, Eric (2010). "Structurally Defined Alternatives and Lexicalizations of XOR". In: *Linguistics and Philosophy* 33.1. Pp. 31–36.

Veltman, Frank (1985). "Logics for Conditionals". PhD thesis. Universiteit van Amsterdam.

Warmbrōd, Ken (1981). "Counterfactuals and Substitution of Equivalent Antecedents". In: *Journal of Philosophical Logic* 10.2. Pp. 267–289.

Wright, Georg Henrik von (1968). *An Essay on Deontic Logic and the Theory of Action*. Amsterdam: North-Holland Publishing Company.

Zimmermann, Thomas Ede (2000). "Free Choice Disjunction and Epistemic Possibility". In: *Natural Language Semantics* 8. Pp. 255–290.