September 2022

The *Only*-Implicature Generalization and its relevance for theories of pragmatics

Daniel Asherov, Danny Fox, & Roni Katzir

Abstract: The literature on scalar implicatures (SIs) varies in its views on the division of labor between grammar and general reasoning in the derivation of SIs. According to the *grammatical approach*, the SIs of a given sentence are logical entailments of particular parses of that sentence – specifically, parses with a silent exhaustivity operator, notated as *Exh*, a covert counterpart of 'only'. According to a competing, pragmatic view, there is no need for anything like *Exh* in the grammar; rather, the dynamics of conversation suffice to derive SIs. The present squib contributes to a comparative evaluation of the two views. We will consider the pragmatic *Exh*-free approach in the context of two frameworks. The first is the *neo-Gricean* framework, in which an utterance leads to the negation of alternatives that are contextually more informative. The second is that of *iterated rationality models* (IRMs). According to this prominent and more recent approach, SIs arise from iterated steps of reasoning by discourse participants about the goals and means available to other discourse participants. We will use a paradigm inspired by one-shot reference games (Rosenberg & Cohen 1964, Frank & Goodman 2012) to argue that SIs track the entailments of overt 'only', thus providing further support to the *Only*-Implicature Generalization (Fox 2004, Fox & Hackl 2006). Our findings support the grammatical approach, which takes the *Only*-Implicature Generalization at face value, but are problematic both for the neo-Gricean approach and for IRMs that are able to derive SIs from non-exhaustified representations.

1. Introduction

A listener who hears (1a) is likely to infer that (1b) is false. This inference is interesting because there are well known arguments that (1a) has a meaning that does not entail the negation of (1b). Practically all responses to these arguments share the assumption that the inference comes about from a process that takes into account not just the basic linguistic expression in (1a) but also alternative expressions. This process leads to what is called a *scalar implicature* (SI). (1c) states the *strengthened meaning* of (1a) – the conjunction of (1a) with its SI that (1b) is false.

(1) a. John ate *some* of the bananas.
b. John ate *all* of the bananas.
c. John ate *some but not all* of the bananas.

In this paper we will compare three different approaches to the derivation of SIs. We do so in light of the *Only*-Implicature Generalization (OIG; Fox 2004, Fox & Hackl 2006), which says that the SIs of a given sentence are the same as the inferences drawn from a variant of that sentence that uses the overt exhaustivity operator 'only' (with an appropriate placement of focus). In (1a), for example, the negation of the alternative (1b) and the strengthened meaning arrived at in (1c) are the same as what we see by using 'only' and placing focus on 'some': "John only ate SOME of the bananas."[1]

The first approach to SIs that we will consider, the so-called *grammatical* approach (see Chierchia, Fox, & Spector 2012, Chierchia 2006, Fox 2007, Fox and Hackl 2006),[2] takes the OIG at face value. According to this approach, the similarity between SIs and 'only' is no accident: it arises due to a silent exhaustivity operator (notated as *Exh*), a covert counterpart of the overt alternative-sensitive operator 'only'. The SI of (1a), for example, arises because, in addition to the parse that does not entail the negation of (1b), there is another parse that does (*Exh* [John ate some of the bananas]).

According to a competing, pragmatic view, there is no need for anything like *Exh* in the grammar; rather, the dynamics of conversation suffice to derive SIs. We will consider two such approaches.[3]

One pragmatic framework that we will consider and that does not rely on *Exh* is the *neo-Gricean* approach (NG; see Horn 1972, Gazdar 1979, and Sauerland 2004, among others), according to which the SIs of an utterance arise from pragmatic reasoning about more informative alternatives. In (1a), for example, the hearer might reason about the more informative alternative (1b) and conclude (with the aid of certain auxiliary assumptions) that, since this alternative was not uttered, it is false. This gives rise to the relevant SI and to the strengthened meaning in (1c).[4]

The second pragmatic framework that we will consider and that does not rely on *Exh* is that of *iterated rationality models* (IRMs; Benz 2006, Benz & van Rooij 2007, Bergen & Goodman 2015, Frank & Goodman 2012, Frank et al. 2016, Franke 2009, 2011, Goodman & Stuhlmüller 2013, Rothschild 2013, Scontras et al. 2018, among others). According to this prominent approach, SIs arise from iterated steps of reasoning by discourse participants about the goals and means available to other discourse participants. In the case of (1a), an IRM might start from the observation that if the relevant alternatives are (1a) and (1b) and if the relevant states of the world are one in which John ate some but not all of the bananas and another in which he ate all of them, then (1b) is a perfect message for the latter state, which in turn leaves (1a) as the only message for the former state. We will discuss IRMs in greater detail below.

[1] The division of labor between assertion and presupposition in SIs and with overt 'only' is different. This distinction will not play a role in what follows.

[2] The approach has roots in earlier work, notably Groenendijk & Stokhof 1984, Krifka 1995, Landman 2000, and Chierchia 2004.

[3] While the two pragmatic approaches that we will consider do not rely on *Exh* they are in principle compatible with it, a possibility that has been explored in recent literature and that we turn to in section 8.

[4] NG was argued against based on a variety of considerations (see Fox 2007, 2014, Magri 2009, 2011, Chierchia, Fox, & Spector 2012, and Rothschild 2013, among others). We will set aside these arguments in our evaluation of NG below.

In the case of (1), all three approaches make the same correct prediction, in line with the OIG, that an utterance of (1a) will give rise to the inference that (1b) is false and to the strengthened meaning in (1c). In other cases, as we show below, the predictions of the three approaches diverge. The grammatical approach, by design, is committed to predicting SIs that track the inferences of overt ‘only’. NG and IRMs, on the other hand, have the means of deriving strengthenings that are different from those of ‘only’. Across a range of scenarios, which we discuss in sections 2 – 6, we will show that this ability of NGs and IRMs to deviate from the OIG is problematic: empirically, the OIG holds, which is correctly predicted by the grammatical approach but remains a challenge for both NG and IRMs. To test for strengthening we will focus primarily on definite noun phrases, where embedded strengthenings interact with the presuppositions of the definite article. Such embedded strengthenings are predicted to arise on the grammatical approach but have been taken to be problematic for pragmatic approaches such as NG and for IRMs (see Cohen 1971, Landman 2000, Chierchia 2004, among others). For the purposes of this discussion, we will follow Horn (1985, 1989) in assuming that pragmatic approaches can achieve strengthening in embedded positions by reasoning about embedded constituents as if they were matrix utterances – so called, meta-linguistic interpretation. While much of our discussion concerns embedded strengthenings, where the definite article makes the acceptability judgments under consideration particularly clear, we will point out in section 7 that a very similar pattern obtains also in matrix positions.

In order to keep the presentation simple we will start with considering versions of NG and IRMs that do not have access to *Exh*. We first look at a scenario which is like (1) in that all three approaches converge in their predictions. We will then consider various modifications of this scenario where an SI is not predicted by the grammatical approach but is predicted by various versions of the IRM or of NG. While the grammatical approach and NG can each be treated as a single proposal, there is a large variety of proposals that fall under the IRM framework. With the goal of making the central ingredients of the IRM approach and the conceivable moves within the framework transparent, we decided to make use of the very simple IRMs of Fox & Katzir (2021). Though these IRMs appear to be very different from the more elaborate systems found in the literature, we think that they are based on core assumptions appealed to in pragmatic accounts of SIs (and hence are particularly meaningful when the challenge at hand is one of overgeneration). The relevant assumptions are the following: (a) that interlocutors share common belief in the purpose of the communicative act (to convey a state of the world, or an epistemic state of the speaker, the set of *states*), (b) that the speaker is restricted in her action to the choice among a designated set of formal alternatives (the set of available *messages*), and (c) that the speaker and hearer are rational agents that can compute the consequences of everything that is common belief.[5] As we note in section 8, however, the challenge to the IRM approach is quite general, and in the appendix we show how our scenarios challenge other IRMs that have been proposed in the literature, such as the Rational Speech Act

[5] From these assumptions it follows that if a particular message is unique in identifying a speaker-state, it would be used and if it’s not used, this is not the speaker's state. We’ve learned over time that not everybody agrees with this evaluation of the IRMs in Fox and Katzir, hence the discussion in section 8 and the appendix.

model of Frank & Goodman (2012), the Iterated Best Response model of Franke (2009, 2011), and Rothschild's (2013) derivation of SIs through iterated elimination of weakly dominated strategies. We will also consider what happens if NG and IRMs are enriched so as to have access to *Exh*, as has been done in recent IRM work (Champollion et al. 2019, Franke & Bergen 2020, Cremers et al. 2022). As we will show, access to *Exh* does not automatically resolve the problem for pragmatic approaches. Our conclusion will be that any theory of SIs must conform with the OIG and that currently the only one that is guaranteed to do so is the grammatical approach.

2. Baseline: Overlapping predictions of Exh and Exh-free NG and IRMs

Consider Scenario A in FIGURE 1 (modeled after Stiller et al. 2011, 2015 and Vogel et al. 2014).[6] Given Scenario A, an utterance of (2a) is acceptable and can be used as a request to pick crate II.[7] The acceptability of (2a) is surprising in light of the presuppositions of the definite article. In particular, the uniqueness presupposition of the definite article requires that there be no more than one salient individual in the extension of the NP 'crate with a banana', but in Scenario A there are two such individuals (crate II and III). The acceptability of (2a) can be explained if the NP can be strengthened to mean crate with a banana and with no apple, which has only crate II in its extension. This is indeed what the OIG predicts: 'crate with a banana' should have the same inferences as 'crate with only a banana'. (2a) as a whole, then, should have a meaning akin to (2b).

(2) a. Pick the crate *with a banana*.
    b. Pick the crate *with a banana and with no apple*.

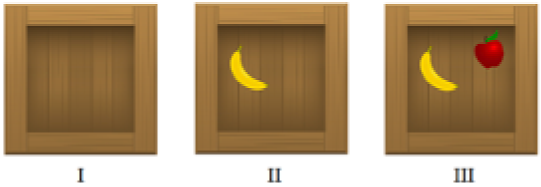


FIGURE 1: SCENARIO A

[6] The scenarios here are based on one-shot reference games (Rosenberg & Cohen 1964, Frank & Goodman 2012), where participants are given a simple message and asked to choose the most appropriate image. Our focus here is different: we are interested in whether an utterance is acceptable in a given scenario and not in what participants choose. While we are not interested in participants' choice per se, we note that the results observed in the experiments in the literature are in line with what can be expected given the acceptability judgments. In particular, although both state II and state III contain a banana, listeners seem to have the intuition that the expression 'banana' refers to state II. Stiller et al.'s participants chose state II 96% of the time. This result was replicated in an experiment by Vogel et al. (2014), who gave only one trial per participant; 75% of their participants chose state II.

[7] Some speakers judge (2a) as less than perfect. What matters for our purposes is not the absolute acceptability of (2a) in this scenario but rather its relative acceptability here and in the scenarios below, where (2a) becomes clearly unacceptable.

In the simple case of Scenario A, taking the set of alternatives to ‘banana’ to consist of single-word fruit names, all three approaches correctly predict the acceptability of (2a). For the grammatical approach this prediction follows the logic of the OIG: if (2a) is parsed as [pick the crate *Exh* with a banana], the sentence will mean the same as (2b), with an NP that has just crate II in its extension.[8] This satisfies the presuppositions of the definite article.

For NG, the reasoning follows informativity, assuming something like Horn’s (1985, 1989) mechanism of meta-linguistic interpretation mentioned above is available. The alternative ‘crate with an apple’ is more informative than ‘crate with a banana’ (since the extension of the former, {crate II}, is a proper subset of the extension of the latter, {crate II, crate III}). So since this more informative alternative was not used, the hearer may conclude (subject to various auxiliary assumptions) that it is false. This leads to the strengthening of the predicate ‘crate with a banana’ to mean crate with a banana and no other fruit – a predicate that has {crate II} as its denotation, a singleton set as required by the uniqueness presupposition of the definite article.

Finally, IRMs derive a similar result. Recall that in this approach, listeners take into account the strategies available to the speaker for achieving a communicative goal. A strategy involves the selection of an utterance from a set of available alternatives, often referred to within the IRM world as *messages*, and the communicative goal, at least in our case, is conveying an intended state. Speakers, in turn, take into account the fact that listeners can use this type of reasoning. Each model in the IRM family implements this idea slightly differently. For now, as mentioned in the introduction, we present a simplified algorithm from Fox & Katzir (2021), in which the core idea is transparent. We consider several modifications of this basic algorithm as we go along and return to the broader range of IRMs in the literature in section 8 and in the appendix. The basic algorithm in our IRM involves iterations in which messages and states are paired together according to an identification criterion and peeled off. The identification criterion we start with, *Semantic State Identification*, is stated in (3).

> **(3) Semantic State Identification:** Given a set of messages *M* and a set of states *T*, a message identifies a state if it is true in that state and there is no other state in which it is true.

In the case of Scenario A, this reasoning proceeds as follows:[9]

[8] Since *Exh* differs from ‘only’ in being silent, there are other conceivable positions in which it might attach within (2a). This will not matter here.

[9] Here and below, messages will be referred to with strings of words between quotation marks, e.g. ‘banana’. States will be labeled by their content with capital initials: in Scenario A state I is labeled EMP (for empty), state II is labeled $\mathbb{B}$ (for banana and no other fruit), and state III is labeled $\mathbb{BA}$ (for banana & apple and no other fruit).

(4) Iterated rationality reasoning about Scenario A (schematic)

Step 1: Although the message 'banana' is initially ambiguous, the message 'apple' is not. Specifically, 'apple' would lead the listener directly to the state $\mathbb{BA}$, since it is inconsistent with the other two states. Thus, if a rational speaker had intended the state $\mathbb{BA}$, they would use the message 'apple'.

Step 2: Given step 1, the speaker who used the message 'banana' did not intend the state $\mathbb{BA}$. Once $\mathbb{BA}$ is eliminated as the intended state, the only remaining option is state $\mathbb{B}$.

As with NG (paired with Horn 1985, 1989's mechanism) and as with the grammatical approach, our basic IRM correctly predicts that the uniqueness presupposition of the definite article is satisfied and that (2a) should be acceptable.

3. A problem for NG and IRMs

We will now modify Scenario A so as to avoid having a crate with a banana and no other fruit. We will see that in such cases "Pick the crate with a banana" is infelicitous, a fact that follows from the OIG and is therefore correctly predicted by the grammatical approach. NG and IRMs, on the other hand, allow for pragmatic reasoning to strengthen "crate with a banana" so as to have only crate II in its extension, which incorrectly predicts the sentence to be acceptable.

Consider Scenario B in FIGURE 2, which is based on Frank & Goodman (2012), Stiller et al. (2011), and Vogel et al. (2014). It differs from Scenario A in the following way: a pear is added to both crate I and crate II. This scenario is important for the evaluation of the OIG and the comparison between the grammatical approach and both NG and IRMs because, unlike Scenario A, here there is no state with *only* a banana.

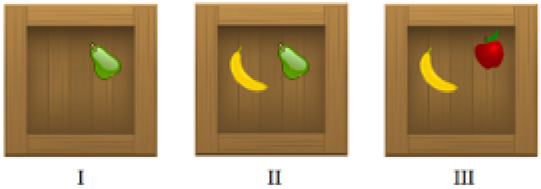


FIGURE 2: SCENARIO B

Differently from Scenario A, in which (2a) was judged as acceptable, in Scenario B the same sentence is judged as unacceptable, presumably because “crate with a banana” is no longer an acceptable way to single out crate II (or to any other individual crate for that matter).[10]

The grammatical approach makes the correct prediction for Scenario B. If (2a) is parsed without *Exh*, then, as in all the scenarios that we consider in this paper, both crate II and crate III are in the extension of “crate with a banana”, so the uniqueness presupposition of the definite article is not satisfied. Differently from Scenario A, however, parsing the sentence with *Exh* is now of little help, as predicted by the OIG. Specifically, the NP [crate *Exh* with a banana], just like [crate with only a banana], means crate with a banana and not with an apple and not with a pear. There is no such crate in this scenario, so the extension of the NP is empty, which leaves the existence presupposition of the definite article unsatisfied. Regardless of whether (2a) is parsed with *Exh*, then, the grammatical approach correctly predicts the sentence to be unacceptable.

Things are less straightforward for NG. As in Scenario A, “crate with a banana” has “crate with an apple” as a strictly more informative alternative: the set of individuals of which “crate with an apple” is true is a strict subset of the set of individual of which “crate with a banana” is true. Note that “crate with a pear” is not more informative than “crate with a banana”; the two are contextually independent, which on common NG assumptions means that “crate with a pear” will not be negated. So “crate with a banana” will be strengthened to mean crate with a banana and not with an apple, which (just as in Scenario A) has exactly one individual — crate II — in its extension. So both the existence and the uniqueness presuppositions of the definite article are satisfied, and (2a) is incorrectly predicted to be acceptable.

Is this a serious problem for NG? It is not that clear. There is a way to construct an NG system that can deal properly with this scenario. Specifically, an NG system could, in principle, be constructed so that it leads to negation of alternatives that are contextually independent of the assertion and not just those that are strictly stronger. In the present case, this means that the candidates for negation will be not just the contextually stronger “crate with an apple” but also the contextually independent “crate with a pear”. Note further that it is impossible to negate both of these alternatives consistently with “crate with a banana”, and that negating either of these alternatives entails that the other holds: if a crate with a banana

[10] Indeed, Vogel et al. (2014) elicited listeners’ judgments in a reference-game experiment and found that only 52% of the participants chose state II (cf. 75% in Scenario A). We wish to reiterate, however, that our focus is on the grammaticality of sentences such as (2a) in the different scenarios and not on how participants choose in corresponding reference games. The correspondence between grammaticality and choice is likely to be complex. For example, when forced to choose in the case of ungrammaticality, as in Scenario B, participants might consider various repairs to the sentence or the images that would make the utterance felicitous. In the present case, one such repair might be to pretend that (2a) has the indefinite ‘a’ instead of the definite ‘the’. We will not attempt to analyze such repairs in this paper. But we do wish to point out that a demonstration that speakers rely on probabilistic reasoning when forced to choose a crate in response to a sentence that is predicted to be ungrammatical by the grammatical theory of SIs (as in Frank & Goodman 2012, Vogel et al. 2014, Stiller et al. 2011, 2015, and Zhou et al. 2022, among others) is not inconsistent with this theory.

is not a crate with an apple then it is a crate with a pear, and vice versa. In such cases it has been argued that neither alternative can be negated (Sauerland 2004), which lets NG predict that "crate with a banana" will not be strengthened, its extension will include both crate II and crate III, and the uniqueness presupposition of the definite article will not be satisfied. This is the correct result.

So there is a way to construct an NG system that would yield the correct result. Still we should ask whether this is an appropriate move within a theory of the decisions that speakers and hearers make in a communicative setting. So we need to ask whether negation of contextually independent alternatives can follow from assumptions about the choices that a speaker would make given the communicative contexts (e.g. assumptions about the maxims that the speaker is following). We think that there are non-trivial issues here since alternatives are not closed under conjunctions. Specifically, a speaker who must choose one alternative among "crate with an apple", "crate with a banana", and "crate with a pear" might be prohibited by the maxims from choosing an alternative that is sub-optimal (not as informative as another alternative). But is it legitimate to demand that a speaker avoid using S just because of an alternative S' that is no better (here, no more informative) than S? This does not seem at all obvious. So we conclude that this scenario is not straightforward for NG. Be that as it may, we will see in section 6 a scenario that is problematic for NG even when we allow this non-trivial choice.

Consider now the pragmatic reasoning on our basic IRM:

> **(5) Iterated rationality reasoning about Scenario B**
>
> Step 1: Although the message 'banana' is initially ambiguous (as is the message 'pear'), the message 'apple' is not. Specifically, 'apple' would lead the listener directly to the state 𝔹𝔸, since it is inconsistent with the other two states. Thus, if a rational speaker had intended the state 𝔹𝔸, they would use the message 'apple'.
>
> Step 2: Given step 1, the speaker who used the message 'banana' did not intend the state 𝔹𝔸. Once 𝔹𝔸 is eliminated as the intended state, the only remaining option is state 𝔹.

It is easy to see that the reasoning in (5) is identical to that in (4), which was used for Scenario A. The IRM algorithm thus incorrectly predicts that listeners will take "crate with a banana" to identify crate II in scenario B as well, which satisfies the uniqueness and existence of the definite article. This, in turn, predicts that (2a) should be acceptable in Scenario B, contrary to fact.

4. Attempting to rescue IRMs through probabilistic identification and blocking iterations

Is there a difference between Scenarios A and B that might be relevant for the workings of an IRM so as to overcome this problem of overgeneration? A potentially relevant difference pertains to the perspective of a speaker that needs to select from among the alternative messages that are true in a given state. In

Scenario A, the message ‘banana’ is the only message that such a speaker can select for crate II but one out of two possible messages for crate III. In Scenario B, on the other hand, ‘banana’ is one out of two possible messages both for crate II and for crate III. This could be stated in probabilistic terms as follows, on the assumption of a *naïve speaker*, who chooses randomly between the messages that are possible given the state they wish to refer to: in Scenario A, ‘banana’ will be uttered with probability 1 for crate II and with probability 0.5 for crate III, while in Scenario B, ‘banana’ will be uttered with probability 0.5 both for crate II and for crate III. Suppose, then, that we changed the identification criterion in a way that can exploit this distinction:[11]

> **(6) Probabilistic State Identification (PSI):** Given a set of messages *M* and a set of states *T*, a message identifies a state if the likelihood that a naïve speaker would use the message to describe that state is higher than for any other state.

In Scenario A, our IRM with PSI correctly finds that the message ‘banana’ refers to state 𝔹. Here, unlike the non-probabilistic State Identification, our IRM discovers this message-cell pairing within one step.

In Scenario B, PSI still arrives at the incorrect result that the message ‘banana’ refers to state II (𝔹ℙ), just like non-probabilistic State Identification:

(7) Iterated rationality reasoning about Scenario B (Probabilistic)

> Step 1: Although the message ‘banana’ does not identify a state, the message ‘apple’ does (as does the message ‘pear’). Specifically, ‘apple’ would lead the listener directly to the state 𝔹𝔸, since P(‘apple’|𝔹𝔸)=0.5>0=P(‘apple’|𝔹ℙ)=P(‘apple’|ℙ). Thus, if a rational speaker who is trying to identify a state with PSI had intended the state 𝔹𝔸, they would use the message ‘apple’. (Similarly, ‘pear’ identifies the state ℙ.)
>
> Step 2: Given step 1, the speaker who used the message ‘banana’ did not intend the state 𝔹𝔸. Once 𝔹𝔸 is eliminated as the intended state, the only remaining option is state 𝔹ℙ.

While PSI still makes the wrong prediction for Scenario B, there is now a distinction between Scenario A and Scenario B: while in Scenario A identification is achieved within one iteration (in Step 1), in Scenario B this is only achieved in the second iteration (only in Step 2).

[11] This probabilistic notion of identification can be motivated in terms of Bayesian reasoning, as is commonly done in the IRM literature. Such motivation, however, leads to the problematic prediction that scalar implicatures should be sensitive to the prior probabilities of states. See Degen et al. (2015), Fox & Katzir (2021), and Cremers et al. (2022) for discussion. Note that whenever *m* identifies *t* by (3) it will also identify *t* by (6) (at least when the set of messages is finite). This means that the move to (6) does not block identification in Scenario B as we discuss below. The only effect of this move on scenarios A and B pertains to the number of steps needed for identification (fewer in A than in B).

Suppose, then, that this difference is responsible for the difference in the acceptability of (2a) in Scenarios A and B. In particular, suppose that humans cannot proceed past the step of iteration needed for identification in Scenario A (past step 1 in our simplified system). If that were the case, the extension of “crate with a banana” would still have crate II (and nothing else) within its extension in Scenario A, which would keep the presuppositions of the definite article satisfied as before. In Scenario B, on the other hand, the restriction to a single step will prevent identification, and both crate II and crate III will be in the extension of “crate with a banana,” which will correctly lead to unacceptability. The goal of the next two sections is to test this idea by constructing scenarios that, on the one hand, have no crate with a banana and no other fruit (so that grammatical exhausitification fails), but, on the other hand, make it possible for IRMs to identify crate II even under the idea that humans are limited to no more than one iteration of PSI.[12] We will find that, even with this limitation, the IRM approach still suffers from an overgeneration problem.

5. The problem arises even within a single iteration

Consider Scenario C (Figure 3), where (2a) is judged unacceptable. Like Scenario B, there are two crates with a banana but no crate with *only* a banana, so the OIG predicts that the sentence will be unacceptable regardless of whether “crate with a banana” is strengthened. The grammatical approach, which tracks the OIG, makes the same correct prediction. In this case, NG also makes the correct prediction, regardless of whether it negates only strictly stronger alternatives than the assertion or also alternatives that are logically independent of it. This is so because “crate with a pear”, “crate with an apple”, and “crate with an orange” are all strictly more informative than “crate with a banana”, but it is impossible to negate all of these more informative alternatives consistently with “crate with a banana”. As mentioned above, this means that no strengthening will take place and that “crate with a banana” will have both crate II and crate III in its extension, which in turn will leave the uniqueness presupposition of the definite article unsatisfied.

While the grammatical approach and NG make the correct prediction for Scenario C, PSI does not, even with the restriction to a single step of identification. This is so since $P(\text{‘banana’}|\mathbb{BP})=0.5>0.33=P(\text{‘banana’}|\mathbb{BAO})$. This should make “crate with a banana” identify crate II within the first iteration, which in turn would satisfy the uniqueness presupposition of the definite article and make (2a) acceptable, contrary to fact.

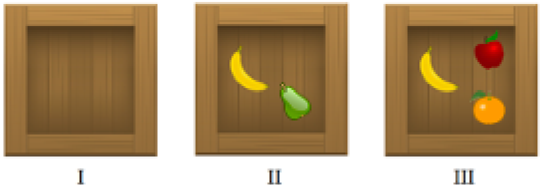


Figure 3: Scenario C

[12] The idea that humans cannot proceed past a first step of PSI clashes directly with the need for multiple steps of PSI in an IRM to account for other inferences such as conjunctive readings of disjunction (see Franke 2009, 2011, van Rooij 2010, and Fox & Katzir 2021). We will set aside this concern for the purposes of the present discussion.

6. Further attempts to avoid the overgeneration problem for IRMs, and a new challenge for NG

We wish to discuss two further potential IRM responses to the overgeneration problem identified above: one response based on the idea that a message that identifies a state must be no worse than other available messages that can serve the same purpose and another response based on the idea that sensitivity to probabilities might be more limited than assumed in PSI. We present minimal variations on scenario C which will demonstrate that neither response can solve the overgeneration problem for IRMs. Both variations will also pose a nontrivial challenge for NG.

6.1. No sub-optimal identifiers

The first response is based on the intuition that a speaker who wants to identify state II in Scenario C above above has a more obvious strategy than using the message ‘banana’, namely using the message ‘pear’ (after all ‘pear’ is true exclusively in state II, while ‘banana’ is true in state II and also in another state). Assuming that this intuition can be grounded in a principle, then we can say that even though ‘banana’ is an identifier for crate II according to PSI, it is a suboptimal message. One can use this observation as the basis for a modification of our IRM so that the problem of overgeneration in Scenario C will not arise. (We thank Anton Benz, p.c., for bringing up this point.) This, together with restricting identification to only one iteration, could give us an IRM story for Scenarios A-C.

But we think that this move will not be helpful. To see why, consider the following minimal variation on Scenario C, where a pear replaces the orange in crate III:

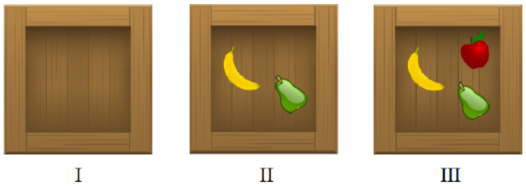


FIGURE 4: SCENARIO C'

In Scenario C', ‘banana’ still identifies crate II using PSI, just like it did in Scenario C. But differently from scenario C, in Scenario C' ‘pear’ is not a better message than ‘banana’ in any conceivable sense. This means that ‘banana’ would identify crate II in Scenario C' by any modification of PSI that avoids identification by suboptimal messages. And yet, the utterance “Pick the crate with a banana” remains as bad in Scenario C' as it was in Scenario C.

Note that as in Scenarios B and C, there are two crates with a banana but no crate with *only* a banana, so the OIG and the grammatical approach make the correct prediction that (2a) will be unacceptable. Differently from Scenario C, NG makes the wrong prediction for Scenario C', regardless of whether only alternatives that are contextually stronger than the assertion are negated or also alternatives that are contextually independent of it. This is so because, as in Scenario B, "crate with a pear" is contextually equivalent to the assertion and so will not be negated in either version of NG, while "crate with an apple" is strictly stronger than the assertion and can be negated on both versions. And when it is negated, "crate with a banana" is strengthened to mean crate with a banana and not with an apple, which satisfies the presuppositions of the definite article. This is a challenge for NG.

6.2 Granularity

The second response to the overgeneration problem for IRMs is based on the idea that our sensitivity to probabilities (at least when computing SIs) is not as fine-grained as assumed in PSI. In Scenario A, where PSI successfully allowed for 'banana' to identify crate II in the first iteration, there was a big difference between a naïve speaker's probability of uttering 'banana' in crates II and III: 1 in crate II and 0.5 in crate III. In Scenarios C and C', where PSI incorrectly allowed for 'banana' to identify crate II (also in the first iteration), the difference between a naïve speaker's probability of uttering 'banana' in the two crates was much smaller: 0.5 in crate II and 0.33 in crate III. To block identification by 'banana' in Scenarios C and C', then, one might propose a less fine-grained version of PSI that allows a message *m* to identify a state *t* only when P(m|t) is sufficiently bigger than P(m|t') for any other cell, and one can imagine various statements of what counts as sufficiently bigger.

We think, however, that no such statement will succeed. To see why, consider the following variant of Scenario C': instead of using only single fruit names in our messages we will now allow also for *conjunctions* of fruit names. In this setting, 'Pick the crate with a banana and a pear' is a good message and is understood as referring to crate II. 'Pick the crate with a banana', on the other hand, is a bad message, just as it was when we considered only single fruit names, a fact that the grammatical approach correctly predicts (and that NG fails to predict, regardless of whether it allows the negation of alternatives that are contextually independent of the assertion) for reasons that are familiar by now. The problem for any attempt to modify PSI based on probability differences by a naïve speaker is that the two messages have the exact same probabilities in crates II and III: 0.33 in crate II and 0.14 in crate III. No way of restricting identification based on probability differences will therefore succeed in allowing 'Pick the crate with a banana and a pear' to identify crate II while preventing 'Pick the crate with a banana' from doing the same.[13]

[13] This property of IRMs persists even if one assigns different costs to different messages, e.g. by message length, as has been suggested for RSA (see Bergen et al. 2016, Scontras et al. 2018).

7. Matrix vs. embedded implicatures

Above we focused on strengthenings in embedded positions. This allowed us to rely on the effect of strengthening on acceptability (given the presuppositions of the definite article), which resulted in a clear pattern. We believe, however, that very similar facts obtain in matrix positions as well.

Consider, for example, the following exchange between two speakers who look at crate II and crate III in Scenario A and B above.

(8) **Speaker A:** What is the difference between crate II and crate III?
Speaker B: In crate II there is a banana. In crate III there is an apple and a banana.

The exchange seems natural in Scenario A but not in Scenario B. We claim that the explanation of this contrast must rely on the OIG. To understand why this is the case, we first note that the utterance by Speaker B, should not be acceptable unless an SI is computed. Without an SI what is said of crate II is that it has a banana in it and possibly other fruit as well. This is also true of crate III and hence does not in any way identify a difference between the two crates. This reasoning is what explains the oddness of (9) and (10) in any context:

(9) **Speaker A:** What is the difference between crate II and crate III?
Speaker B: #In crate II there is an apple or a banana. In crate III there is a banana.

(10) **Speaker A:** What is the difference between Mary and John?
Speaker B: #Mary lives in France. John lives in Paris.

8. Other moves within pragmatic theories

We saw that, across a range of scenarios, strengthening is possible exactly when ‘only’ yields strengthening. Specifically, we considered scenarios in which there are two crates with a banana, so without strengthening the uniqueness presupposition of the definite article in “the crate with a banana” is not satisfied. In such scenarios, “Pick the crate with a banana” is judged acceptable exactly when “crate with *only* a banana” has just one individual in its extension. We then showed that a similar pattern obtains in matrix positions: strengthening corresponds to the entailments of ‘only’. This is, of course, expected under the OIG and the grammatical approach. On the other hand, it is surprising for theories that allow for pragmatic strengthening, be they of the NG variety or IRMs. We showed this for NG (both with the negation of strictly more informative alternatives and, using Scenario C', with the negation also of logically independent ones) and for several variations on a basic IRM. In the appendix we show that the same holds also for several prominent IRMs from the literature: the Rational Speech Act model (Frank &

Goodman 2012); Iterated Best Response (Franke 2009); and the elimination of weakly dominated strategies (Rothschild 2013). We take this failure of *Exh*-free pragmatic models as further evidence that pragmatic theories cannot replace *Exh*, a conclusion that is very much in line with Champollion et al. (2019), Franke & Bergen (2020), Fox & Katzir (2021), and Cremers et al. (2022). But the scenarios above argue for a stronger conclusion. Even with a grammar that includes *Exh*, a pragmatic model could in principle yield the same problematic strengthened meanings as before, simply by using parses without *Exh* and proceeding as in our discussion above. This would leave such *Exh*-enhanced pragmatic models with the same overgeneration problem as before. If pragmatic models of the NG and IRM kind are to be maintained, then, they must not just work with a grammar that has *Exh* but also be prevented from yielding strengthenings in the scenarios above other than through *Exh*.

If one adopts an NG or IRM approach to model pragmatic interactions, one needs to understand why it is that this model relies on *Exh* for strengthening and cannot also derive strengthening in other ways.

If one were to focus exclusively on the examples of strengthening within the scope of the definite article discussed in sections 2–6, a natural explanation for the dependence on *Exh* could be that pragmatic reasoning does not, after all, have the means of strengthening in embedded positions, *contra* Horn (1985, 1989). If that were the case, a parse with *Exh* would be the only way of strengthening in the environments discussed in the scenarios above (and more generally the only way of deriving embedded implicatures). As we showed in section 7, however, a very similar pattern obtains also in matrix position.

A different possibility is that, as briefly considered above, humans happen to be limited to a single iteration of an IRM for some reason. As we have seen, this single iteration cannot be probabilistic, since that would incorrectly lead to identification in Scenarios C and C', as discussed above. (Note that in a non-probabilistic model limited to one iteration, identification in Scenario A would be possible only with '*Exh*(banana)'.)

Yet another possibility relates to a possible distinction between formal alternatives used by *Exh* and *only* in grammar on the one hand and the alternatives that people consider in pragmatic reasoning (when reasoning about each other's communicative goals and strategies) on the other hand. The alternatives used by *Exh* and *only* in grammar have been argued to be rather restricted (Fox & Katzir 2011, Trinh & Haida 2015). In our earlier presentation we assumed that the set of messages in a pragmatic model (NG or IRM) can be similarly restricted. This assumption, which is shared with earlier work such as Horn (1972), Gazdar (1979), Frank & Goodman (2012), Stiller et al. (2011, 2015), and Vogel et al. (2014), can be questioned. It is possible that during pragmatic reasoning humans actually have access, *by necessity*, to all the sentences in their language as potential messages. (See Fox 2007, 2014, Fox & Katzir 2011 for conceptual and empirical motivation.) In this case, there will always be an exact alternative which can be used by NG and which will lead to identification already in the first iteration of an IRM, with the consequence that no further iterations could lead to any enrichment of meaning.

Other possibilities can presumably be imagined as well. Whatever the correct explanation, and whatever is ultimately concluded about the suitability of NG or of IRMs as models of pragmatics, we are left with the conclusion that the mechanism for strengthening must obey the OIG, i.e., that when *Exh* does not yield a strengthening – as in Scenarios B, C, and C' above – there is no other strengthening mechanism that can bypass it.

References

Benz, Anton and Robert Van Rooij (2007), Optimal assertions, and what they implicate. A uniform game theoretic approach. *Topoi* **26(1)**; 63-78.

Benz, Anton (2006), Utility and Relevance of Answers. In Benz A., Jäger G., van Rooij R. (eds). *Game Theory and Pragmatics*. Palgrave Macmillan, London. 195-219.

Bergen, Leon and Noah D. Goodman (2015), The strategic use of noise in pragmatic reasoning. *Topics in cognitive science* **7(2)**; 336-350.

Bergen, Leon, Roger Levy, and Noah D. Goodman (2016), Pragmatic reasoning through semantic inference. *Semantics and Pragmatics* **9(20)**.

Champollion, Lucas, Anna Alsop, and Ioana Grosu (2019), Free choice disjunction as a rational speech act. *Proceedings of SALT* **29**; 238-257.

Chierchia, Gennaro (2004), Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface. In A. Belletti (ed.), *Structures and Beyond 3*. Oxford: Oxford University Press. 39–103 .

Chierchia, Gennaro, Danny Fox, and Benjamin Spector (2012), Scalar implicature as a grammatical phenomenon. In P. Portner, C. Maienborn, and K. von Heusinger (eds.), *Semantics: An international handbook of natural language meaning*, volume 3. Mouton de Gruyter. 2297–2233.

Cohen, L. Jonathan (1971), Some remarks on Grice's views about the logical particles of natural language. In Bar-Hillel, Y., editor, Pragmatics of Natural Language. Reidel. 50–68.

Cremers, Alexandre, Ethan Wilcox, and Benjamin Spector (2022), Exhaustivity and anti-exhaustivity in the RSA framework: Testing the effect of prior beliefs. Ms.

Degen, Judith, Michael Henry Tessler, and Noah D. Goodman (2015), Wonky worlds: Listeners revise world knowledge when utterances are odd. In Noelle, D. C., Dale, R., Warlaumont, A. S., Yoshimini, J., Matlock, T., Jennings, C. J., and Maglio, P. P., (eds.) *Proceedings of the 37th Annual Conference of the Cognitive Science Society*. 548–553.

Fox, Danny (2004), Implicatures and exhaustivity. Handout for a seminar given at USC.

Fox, Danny (2007), Free choice disjunction and the theory of scalar implicatures. In U. Sauerland and P. Stateva (eds.), *Presupposition and implicature in compositional semantics*. Palgrave-Macmillan. 71–120.

Fox, Danny (2014), Canceling the Maxim of Quantity: Another challenge for a Gricean theory of scalar implicatures. *Semantics and Pragmatics* **7**; 5-1.

Fox, Danny and Martin Hackl (2006), The universal density of measurement. *Linguistics and Philosophy* **29**; 537–586.

Fox, Danny and Roni Katzir (2011), On the characterization of alternatives. *Natural Language Semantics* **19**; 87–107.

Fox, Danny and Roni Katzir (2021), Notes on iterated rationality models of scalar implicatures. *Journal of Semantics* **38(4)**; 571–600.

Frank, Michael C., and Noah D. Goodman (2012), Predicting pragmatic reasoning in language games. *Science* **336**; 998.

Frank, Michael C., Andrés G. Emilsson, Benjamin Peloquin, Noah D. Goodman, and Christopher Potts (2016), *Rational speech act models of pragmatic reasoning in reference games*. Unpublished Ms.

Franke, Michael (2009), *Signal to act: Game theory in pragmatics*. Amsterdam: Institute for Logic, Language and Computation.

Franke, Michael (2011), Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics* **4**; 1-82.

Franke, Michael and Leon Bergen (2020), Theory-driven statistical modeling for se- mantics and pragmatics: A case study on grammatically generated implicature readings. *Language* **96(2)**; e77–e96.

Gazdar, Gerald (1979), *Pragmatics: Implicature, presupposition, and logical form*. New York: Academic Press.

Goodman, Noah. D., and Andreas Stuhlmüller (2013), Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science* **5(1)**; 173-184.

Groenendijk, Jeroen Antonius Gerardus and Martin Johan Bastiaan Stokhof (1984), *Studies in the Semantics of Questions and the Pragmatics of Answers*. PhD thesis, Universiteit van Amsterdam, Amsterdam.

Horn, Laurence R (1972), *On the semantic properties of logical operators in English*. UCLA PhD thesis.

Horn, Laurence R (1985), Metalinguistic negation and pragmatic ambiguity. *Language* **61(1)**; 121–174.

Horn, Laurence R (1989), *A natural history of negation*. Chicago, IL: University of Chicago Press.

Krifka, Manfred (1995), The semantics and pragmatics of polarity items. *Linguistic Analysis* **25**; 1–49.

Landman, Fred (2000), *Events and Plurality: The Jerusalem Lectures*. Kluwer Academic Publishers, Dordrecht.

Magri, Giorgio (2009), A theory of individual-level predicates based on blind mandatory scalar implicatures. *Natural Language Semantics* **17(3)**; 245–297.

Magri, Giorgio (2011), Another argument for embedded scalar implicatures based on oddness in downward entailing environments. *Semantics and Pragmatics* **4(6)**; 1–51.

Rosenberg, Seymour, and Cohen, Bertram D. (1964), Speakers' and listeners' processes in a word-communication task. *Science* **145(3637)**; 1201-1203.

Rothschild, Daniel (2013), Game theory and scalar implicatures. *Philosophical Perspectives* **27**; 438-478.

Sauerland, Uli (2004), Scalar implicatures in complex sentences. *Linguistics and Philosophy* **27(3)**; 367–391.

Scontras, Gregory, Michael Henry Tessler, and Michael Franke (2018), Probabilistic language understanding: An introduction to the Rational Speech Act framework. Retrieved 2021-7-8 from https://www.problang.org.

Stiller, Alex, Noah Goodman, and Michael C. Frank (2011), Ad-hoc scalar implicature in adults and children. *Proceedings of the 33rd annual meeting of the Cognitive Science Society*; 2134-2139.

Stiller, Alex, Noah Goodman, and Michael C. Frank (2015), Ad-hoc implicature in preschool children. *Language Learning and Development* **11(2)**; 176-190.

Trinh, Tue and Haida, Andreas (2015), Constraining the derivation of alternatives. *Natural Language Semantics* **23(4)**; 249–270.

van Rooij, Robert (2010), Conjunctive interpretation of disjunction. *Semantics and Pragmatics* **3**; 1–28.

Vogel, Adam, Andreas G. Emilsson, Michael C. Frank, Dan Jurafsky, and Christopher Potts (2014), Learning to reason pragmatically with cognitive limitations. *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*; 3055–3060.

Zhou, Irene, Jennifer Hu, Roger Levy, and Noga Zaslavsky (2022), Teasing apart models of pragmatics using optimal reference game design. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Appendix: relating PSI to more complex IRMs

In this appendix we compare the predictions of our simplified IRM based on Probabilistic State Identification (PSI, akin to Cell Identification in Fox & Katzir 2021) with more complex models proposed in the IRM literature, namely the Rational Speech Act model (Frank & Goodman 2012, Goodman & Stuhlmüller 2013, Bergen & Goodman 2015, Frank et al. 2016, Scontras et al. 2018, and others), Iterated Best Response (Franke 2009, 2011), and iterative elimination of weakly dominated strategies (Rothschild 2013). We show how the predictions of PSI presented in this paper extend to these IRMs. The demonstrations that follow are not meant to be read as an introduction to the respective IRMs, but we outline the main mechanics of the systems to illustrate how they derive inferences in the context of the scenarios discussed in this paper.

1. Rational Speech Act

We start with the Rational Speech Act model (RSA), which received much attention in the IRM literature over the past decade. The basic principle in RSA is that discourse participants (speakers and listeners) reason about each other's available communicative strategies and their respective probabilities. Listeners generate a probability distribution over states given each available message, and speakers, in turn, use the listeners' probability distributions to determine the probabilities they assign to messages given states. Listeners then make an analogous inference based on the speaker's output, and so on. This recursive computation uses intermediate representations of model-internal speakers and listeners until it arrives at a final outcome.

The update in probabilities at each step is calculated using Bayesian inference (in 1), such that the probabilities which the listener assigns to each of the states given each message ($P_L(t|m)$) are proportional to the probabilities assigned by the speaker to each available message given each of these states ($P_S(m|t)$) multiplied by the prior of the state ($P(t)$). The model is initiated with a *literal listener* that assumes that for each state, every message that is true in that state is equally likely.

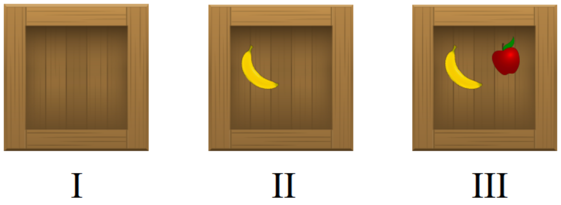
$$(1)\quad P_L(t|m) = \frac{P_S(m|t)P(t)}{\sum_{t' \in T} P_S(m|t')P(t')}$$

In what follows we will assume that the states have equal prior probabilities. In our scenarios, which always have exactly three available states, each state will have a prior of ⅓. In addition to simplifying the discussion, the assumption of flat priors is sensible given the goal of the current exercise, which is to show that the RSA overgenerates implicatures. By assuming flat priors we are giving the RSA the best starting point, because this derives a tie between states II and III in the beginning of the computation in

Scenarios B – C'. We will see that despite this help the RSA still avoids a tie and predicts a strengthening in favor of state II. See Fox & Katzir (2021) and Cremers et al. (2022) for further discussion of priors in IRMs and for problems arising in these models from assuming nonflat priors.

RSA can model listeners who are only approximately rational by limiting the number of speaker-listener iterations and by adjusting a free parameter whose setting can reduce the gained advantage of the optimal message(s). We will return to these limitations later on, and will consider the consequences that they have on the parallels between PSI and RSA.

Scenario A: We start by working through the RSA computation for Scenario A (FIGURE 1). The algorithm starts with a strictly literal listener (L_0) who distributes equally the probabilities in each state among the messages true in that state. In FIGURE 1, the matrix columns correspond to states and the matrix rows correspond to messages. Since a listener's task is to pick among states for a given message, the sum of all probabilities in a single row is 1. In Scenario A, L_0 assigns probability 1 to state 𝔹𝔸 given the message 'apple' because this is the only state in which the message 'apple' is true. Given the message 'banana', the literal listener assigns a probability of 0.5 to each of the states in which it is true, since there are exactly two such states (𝔹 and 𝔹𝔸). The probabilities of all other message-state pairs are equal to 0.



L_0

	EMP	𝔹	𝔹𝔸
'apple'	0	0	1
'banana'	0	0.5	0.5

S_1

	EMP	𝔹	𝔹𝔸
'apple'	0	0	0.67
'banana'	0	1	0.33

L_1

	EMP	𝔹	𝔹𝔸
'apple'	0	0	1
'banana'	0	0.75	0.25

S_2

	EMP	𝔹	𝔹𝔸
'apple'	0	0	0.8
'banana'	0	1	0.2

L_2

	EMP	𝔹	𝔹𝔸
'apple'	0	0	1
'banana'	0	0.83	0.17

FIGURE 1: Scenario A in RSA

Next, the algorithm computes the best strategy available to the speaker S_1 considering L_0's assumptions. The probability that S_1 assigns to a certain message given a certain state is derived by dividing the corresponding cell in L_0 by the sum of the probabilities in its column. For example, the probability that S_1 assigns to the message 'banana' given state 𝔹 is calculated by dividing the probability of state 𝔹 given 'banana' in L_0 ,which is 0.5, by the sum of probabilities of its column, which is also 0.5, since it is the only non-zero cell in this column (because no other message is true in 𝔹). This computation is demonstrated in (2) with the formula in (1) applied to the speaker, assuming, for now, uniform priors for all messages.[1]

$$(2)\quad P_{S1}('banana'|\mathbb{B}) = \frac{P_{L0}(\mathbb{B}|'banana')P('banana')}{P_{L0}(\mathbb{B}|'banana')P('banana') + P_{L0}(\mathbb{B}|'apple')P('apple')} = \frac{0.5*0.5}{0.5*0.5 + 0*0.5} = 1$$

Formula (3) shows the computation of the probability that S_1 assigns to the message 'banana' given 𝔹𝔸, which differs from 𝔹 in that there is another message that can describe it, 'apple'. This probability is calculated by dividing the probability of the state 𝔹𝔸 given 'banana' in L_0, 0.5, by the sum of probabilities of its column, which is equal to 1.5 . The resulting probability assigned to S_1 to 'banana' given 𝔹𝔸 is 0.33, demonstrated in (3). The corresponding probability of 'apple' is 0.67, which means that S_1 is more likely to utter the message 'apple' (0.67) than the message 'banana' (0.33) given 𝔹𝔸.

$$(3)\quad P_{S1}('banana'|\mathbb{BA}) = \frac{P_{L0}(\mathbb{BA}|'banana')P('banana')}{P_{L0}(\mathbb{BA}|'banana')P('banana') + P_{L0}(\mathbb{BA}|'apple')P('apple')} = \frac{0.5*0.5}{0.5*0.5+ 1*0.5} = 0.33$$

Given S_1, a "pragmatic listener" L_1 can use the same kind of Bayesian reasoning to revise the probability that should be assigned to each state given each message. To find the probability of state 𝔹 given 'banana', the algorithm divides the probability that S_1 assigned to state 𝔹 by the sum of the probabilities that S_1 assigned to each state. The computations for 𝔹 and 𝔹𝔸 given 'banana' in L_1 are shown in (4) and (5), respectively (recall that the probabilities of state ø is equal to 0 for both messages). The outcome of the computation in S_1 and L_1 is that the pragmatic listener L_1 is more inclined to interpret 'banana' as referring to state 𝔹 (0.75) rather than to 𝔹𝔸 (0.25). In the following speaker-listener iteration, the probability that L_2 assigns to 𝔹 grows from 0.75 (in L_1) to 0.83.

$$(4)\quad P_{L1}(\mathbb{B}|'banana') = \frac{P_{S1}('banana'|\mathbb{B})P(\mathbb{B})}{P_{S1}('banana'|\mathbb{B})P(\mathbb{B}) + P_{S1}('banana'|\mathbb{BA})P(\mathbb{BA}) + P_{S1}('banana'|ø)P(ø)} = \frac{1*0.3}{1*0.3 + 0.3*0.3+0*0.3} = 0.75$$

$$(5)\quad P_{L1}(\mathbb{BA}|'banana') = \frac{P_{S1}('banana'|\mathbb{BA})P(\mathbb{BA})}{P_{S1}('banana'|\mathbb{B})P(\mathbb{B}) + P_{S1}('banana'|\mathbb{BA})P(\mathbb{BA}) + P_{S1}('banana'|ø)P(ø)} = \frac{0.3*0.3}{1*0.3 + 0.3*03+0*0.3} = 0.25$$

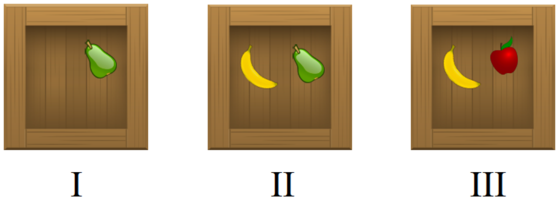
We now can see that the simplified IRM with Probabilistic State Identification, repeated in (6), is an approximation of the listener's inference in RSA and shares its underlying rationale. The listener (L_1) infers the probabilities of each state given each message based on the probabilities that the speaker (S_1)

[1] This assumption will be discussed below in the context of Scenario C'.

assigns to each message given each state. Both models eventually arrive at the outcome that ‘banana’ will identify state 𝔹. This is the desired result (§2 in the main text).

(6) **Probabilistic State Identification (PSI):** Given a set of messages *M* and a set of states *T*, a message identifies a state if the likelihood that the speaker would use the message to describe that state is higher than for any other state.

Scenarios B and C: The computations for Scenarios B (FIGURE 2) and C (FIGURE 3) are similar, as well as their prediction that ‘banana’ will identify their respective state II. In Scenario B, the speaker S_1 assigns a higher probability to the message ‘banana’ given state 𝔹 (0.5) than given state 𝔹𝔸 (0.33); accordingly, by the formula in (1), L_2 arrives at a higher probability for state 𝔹 given ‘banana’ (0.6) than for state 𝔹𝔸 given ‘banana’ (0.4). Here, too, both PSI and RSA eventually arrive at the outcome that ‘banana’ will identify state 𝔹 in Scenario B, and similarly ‘banana’ will identify state 𝔹ℙ in Scenario C. Note that unlike in Scenario A, these results deviate from the *Only*-Implicature Generalization (see discussion in §3-5 in the main paper).



L_0

	ℙ	𝔹ℙ	𝔹𝔸
‘apple’	0	0	1
‘banana’	0	0.5	0.5
‘pear’	0.5	0.5	0

S_1

	ℙ	𝔹ℙ	𝔹𝔸
‘apple’	0	0	0.67
‘banana’	0	0.5	0.33
‘pear’	1	0.5	0

L_1

	ℙ	𝔹ℙ	𝔹𝔸
‘apple’	0	0	1
‘banana’	0	0.6	0.4
‘pear’	0.67	0.33	0

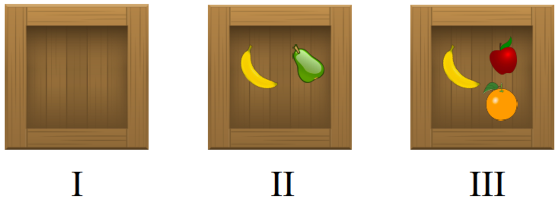
S_2

	ℙ	𝔹ℙ	𝔹𝔸
‘apple’	0	0	0.71
‘banana’	0	0.65	0.29
‘pear’	1	0.35	0

L_2

	ℙ	𝔹ℙ	𝔹𝔸
‘apple’	0	0	1
‘banana’	0	0.69	0.31
‘pear’	0.74	0.26	0

FIGURE 2: Scenario B in RSA



	L_0			S_1			L_1			S_2			L_2		
	EMP	𝔹ℙ	𝔹𝔸𝕆	EMP	𝔹ℙ	𝔹𝔸𝕆	EMP	𝔹ℙ	𝔹𝔸𝕆	EMP	𝔹ℙ	𝔹𝔸𝕆	EMP	𝔹ℙ	𝔹𝔸𝕆
‘apple’	0	0	1	0	0	0.4	0	0	1	0	0	0.42	0	0	1
‘banana’	0	0.5	0.5	0	0.33	0.2	0	0.62	0.38	0	0.38	0.16	0	0.7	0.3
‘pear’	0	1	0	0	0.66	0	0	1	0	0	0.62	0	0	1	0
‘orange’	0	0	1	0	0	0.4	0	0	1	0	0	0.42	0	0	1

Figure 3: Scenario C in RSA

The RSA model can in principle prevent identification even in cases in which the strengthened interpretation has the highest likelihood (state II in Scenarios A-C) by stopping the model before likelihood differences are maximized, e.g., through a parameter limiting the depth of the recursion. For example, in Scenarios A-C, if the algorithm stops at L_1, under some assumptions, the model assigns a probability of 0.75 to state II given ‘banana’ in Scenario A (state 𝔹), 0.6 in Scenario B (state 𝔹ℙ), and 0.62 in Scenario C (state 𝔹ℙ).[2] Probabilities can be further diminished through another parameter, which moderates the magnitude of the changes in probabilities from one iteration to the next. In RSA this is the *rationality* parameter, sometimes labeled α, which is incorporated into each inference step as in (7).

$$(7)\quad P_L(t|m) = \frac{exp(\alpha\cdot logP_S(m|t))}{\sum_{t'\in T} exp(\alpha\cdot logP_S(m|t'))}$$

With these parameters, RSA can derive the result that 'banana' is more likely to identify state II in Scenario A than in the other two scenarios. Nevertheless, the basic problem with Scenarios B and C remains unsolved: regardless of the settings of the model’s parameters, state II always has the highest probabilities in Scenario B and C, even if this preference is small (Figure 4). We leave open the

[2] Such probabilities derived by RSA have been used to model participants' preferences in reference game experiments (Frank & Goodman 2012, Stiller et al. 2011, Vogel et al. 2014, and others). Here we are concerned with the extension assigned to "crate with a banana" and its consequences for satisfying the presuppositions of the definite article in "Pick the crate with a banana".

possibility that RSA-type parameter settings which reflect human behavior are such that this preference is unnoticed (small number of iterations, small α value), leading to ungrammaticality of "Pick the crate with a banana" in these scenarios.

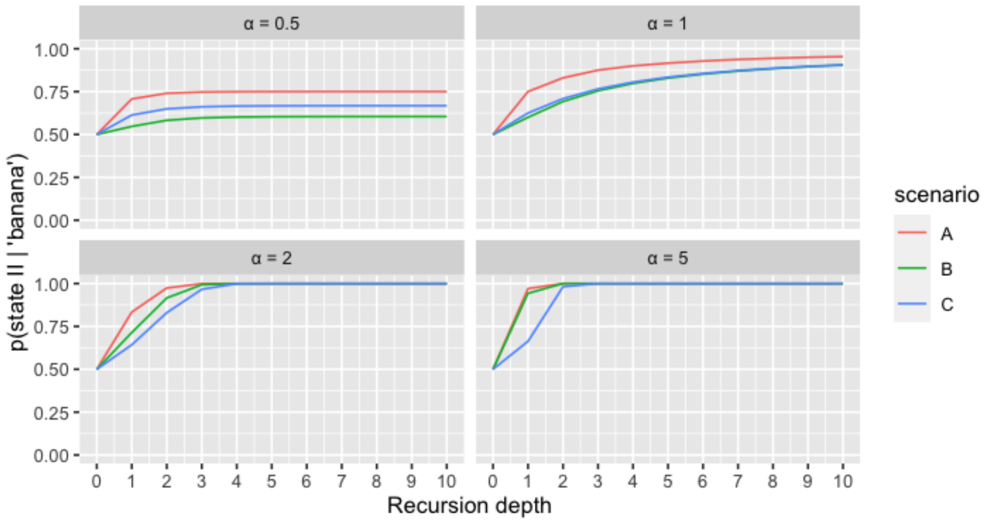
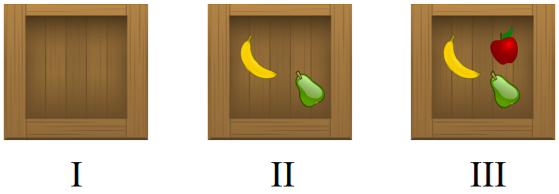


FIGURE 4: Predicted probability of state II in RSA by recursion depth and α value with flat priors

Scenario C': In Scenario C above, although RSA identifies state II as the best interpretation for 'banana', the other possible message in state II, 'pear', is always preferred by the speaker. The reason is that 'pear' is true exclusively in state II, while 'banana' is also true in another state. Scenario C' shows that this solution does not extend to other similar cases. In Scenario C', both 'banana' and 'pear' are true in both states II and III, so that their distribution is identical, and no message is sub-optimal with respect to the other. As FIGURE 5 shows, in this case RSA predicts that both 'banana' and 'pear' would be interpreted as referring to state II.



L_0

	EMP	𝔹ℙ	𝔹𝔸ℙ
‘apple’	0	0	1
‘banana’	0	0.5	0.5
‘pear’	0	0.5	0.5

S_1

	EMP	𝔹ℙ	𝔹𝔸ℙ
‘apple’	0	0	0.5
‘banana’	0	0.5	0.25
‘pear’	0	0.5	0.25

L_1

	EMP	𝔹ℙ	𝔹𝔸ℙ
‘apple’	0	0	1
‘banana’	0	0.66	0.33
‘pear’	0	0.66	0.33

S_2

	EMP	𝔹ℙ	𝔹𝔸ℙ
‘apple’	0	0	0.6
‘banana’	0	0.5	0.2
‘pear’	0	0.5	0.2

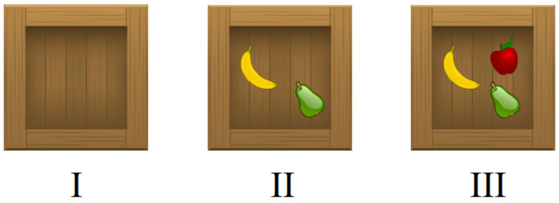
L_2

	EMP	𝔹ℙ	𝔹𝔸ℙ
‘apple’	0	0	1
‘banana’	0	0.71	0.29
‘pear’	0	0.71	0.29

FIGURE 5: Scenario C’ in RSA

Scenario C’* with conjunctive messages*: In §6.2 of the main text we discuss a case in which PSI predicts two messages to pattern identically, but empirically only one of them gives rise to an SI. We consider a version of Scenario C’ in which the speaker and the listener also take into account messages with more than one fruit. The key property of this scenario is that the set of possible messages in state II (𝔹ℙ) is a proper subset of the set of possible messages in state III (𝔹ℙ𝔸). Assuming messages have uniform priors, all messages that are true in state II are more likely in state II than in state III, including both ‘banana’ and ‘banana and pear’ (as well as ‘pear’). Like PSI, the RSA model generates an SI in this scenario, predicting that ‘banana’ and ‘banana and pear’ both identify state II (and also share the same likelihood patterns). This result is at odds with the intuition that ‘banana’ identifies state II while ‘banana and pear’ does not.[3] The relevant derivation in RSA is shown in FIGURE 6 (for concreteness, we assume α=1, as we did in FIGURES 1-3 and 5, as well as uniform priors).

[3] Note that the empirical contrast between ‘banana’ and ‘banana and pear’ cannot be attributed to the fact that the latter is conjunctive and the former is not. To see this, consider a version of Scenario C’ in which states II and III also contain an orange. In this case, the (conjunctive) message ‘banana and pear and orange’ is interpreted as referring to state II, but no similar inference arises given the (also conjunctive) message ‘banana and pear’.



L$_0$

	EMP	𝔹ℙ	𝔹𝔸ℙ
‘apple’	0	0	1
‘banana’	0	0.5	0.5
‘pear’	0	0.5	0.5
‘a & b’	0	0	1
‘a & p’	0	0	1
‘b & p’	0	0.5	0.5
‘a & b & p’	0	0	1

S$_1$

	EMP	𝔹ℙ	𝔹𝔸ℙ
‘apple’	0	0	0.18
‘banana’	0	0.33	0.09
‘pear’	0	0.33	0.09
‘a & b’	0	0	0.18
‘a & p’	0	0	0.18
‘b & p’	0	0.33	0.09
‘a & b & p’	0	0	0.18

L$_1$

	EMP	𝔹ℙ	𝔹𝔸ℙ
‘apple’	0	0	1
‘banana’	0	0.79	0.21
‘pear’	0	0.79	0.21
‘a & b’	0	0	1
‘a & p’	0	0	1
‘b & p’	0	0.79	0.21
‘a & b & p’	0	0	1

S$_2$

	EMP	𝔹ℙ	𝔹𝔸ℙ
‘apple’	0	0	0.22
‘banana’	0	0.33	0.04
‘pear’	0	0.33	0.04
‘a & b’	0	0	0.22
‘a & p’	0	0	0.22
‘b & p’	0	0.33	0.04
‘a & b & p’	0	0	0.22

L$_2$

	EMP	𝔹ℙ	𝔹𝔸ℙ
‘apple’	0	0	1
‘banana’	0	0.9	0.1
‘pear’	0	0.9	0.1
‘a & b’	0	0	1
‘a & p’	0	0	1
‘b & p’	0	0.9	0.1
‘a & b & p’	0	0	1

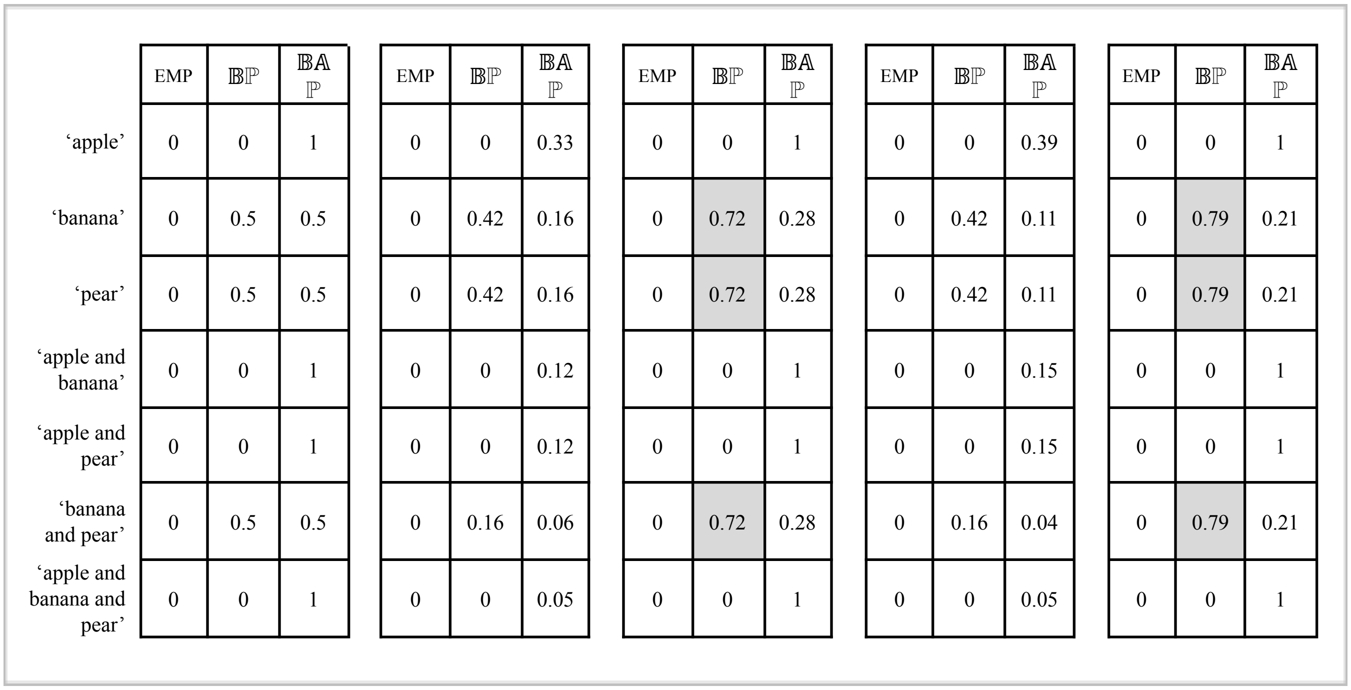
FIGURE 6: Scenario C’ with multi-word messages in RSA

*Gray cells highlight equivalence of probabilities of ‘banana’, ‘pear’, and ‘banana and pear’ given state II.

Is it possible that the source of the asymmetry between the judgment in ‘banana’ and that in ‘banana and pear’ is rooted in a general preference of speakers to use one message over the other? For example, a speaker may be more inclined to utter ‘banana’ rather than ‘banana and pear’ because the latter is longer and phonetically more effortful. This possibility has been modeled in RSA by incorporating the *cost* of a message into the agents’ reasoning (Bergen et al. 2016, Scontras et al. 2018) as in (8), again assuming uniform priors.

(8) $$P_L(t|m) = \frac{exp(\alpha \cdot log(P_S(m|t)) - C(m))}{\sum_{t' \in T} exp(\alpha \cdot log(P_S(m|t')) - C(m))}$$

Such a move does not solve the problem in the case at hand. This is because in Scenario C’, whenever the speaker’s probability to utter ‘banana’ in 𝔹ℙ increases and the probability to utter ‘banana and pear’ in 𝔹ℙ decreases, the corresponding probabilities of the two messages in 𝔹ℙ𝔸 change proportionally. Since the pragmatic listener is sensitive to proportions among probabilities assigned by the speaker, rather than absolute differences, the probability distributions of messages within each state are not affected. For example, let us assume a cost of 1 to each fruit in the message, such that ‘banana’ would have a cost of 1, ‘banana and pear’ a cost of 2, and so forth. The resulting computation is shown in Figure 7. While the probability assigned by the speaker S_1 to ‘banana’ in 𝔹ℙ is higher than that of ‘banana and pear’ in 𝔹ℙ (0.42 and 0.16, respectively), the listener L_1 does not distinguish between the two messages. This is because their relative probability in 𝔹ℙ𝔸 (0.16 and 0.06, respectively) is identical (0.16/0.42=0.06/0.16≈0.368).



	EMP	𝔹ℙ	𝔹𝔸ℙ	EMP	𝔹ℙ	𝔹𝔸ℙ	EMP	𝔹ℙ	𝔹𝔸ℙ	EMP	𝔹ℙ	𝔹𝔸ℙ	EMP	𝔹ℙ	𝔹𝔸ℙ
‘apple’	0	0	1	0	0	0.33	0	0	1	0	0	0.39	0	0	1
‘banana’	0	0.5	0.5	0	0.42	0.16	0	0.72	0.28	0	0.42	0.11	0	0.79	0.21
‘pear’	0	0.5	0.5	0	0.42	0.16	0	0.72	0.28	0	0.42	0.11	0	0.79	0.21
‘apple and banana’	0	0	1	0	0	0.12	0	0	1	0	0	0.15	0	0	1
‘apple and pear’	0	0	1	0	0	0.12	0	0	1	0	0	0.15	0	0	1
‘banana and pear’	0	0.5	0.5	0	0.16	0.06	0	0.72	0.28	0	0.16	0.04	0	0.79	0.21
‘apple and banana and pear’	0	0	1	0	0	0.05	0	0	1	0	0	0.05	0	0	1

Figure 7: Scenario C’ in RSA with cost of 1 per word

*Gray cells highlight equivalence of probabilities of ‘banana’, ‘pear’, and ‘banana and pear’ given state II.

2. Iterated Best Response

Another prominent IRM is the Iterated Best Response (IBR) model proposed in Franke (2009, 2011). Pragmatic reasoning is modeled as an iterative mapping between states to messages and messages to states based on their informativity. We describe how the model derives the prediction for an SI in Scenario B, and provide diagrams for Scenario B and all other scenarios in FIGURE 8 and FIGURE 9.

The algorithm goes as follows, assuming we start with a speaker (see below). In the first step, a speaker S_0 maps states to messages (the messages are labeled by the first letter of the fruit name). We draw connections from each state to all messages which are true in that state. In Scenario B, 5 connections will be drawn in step S_0 – one from state I, which connects to ‘pear’, two from state II, which connects to ‘pear’ and ‘banana’, and two from state III, which connects to ‘banana’ and ‘apple’.

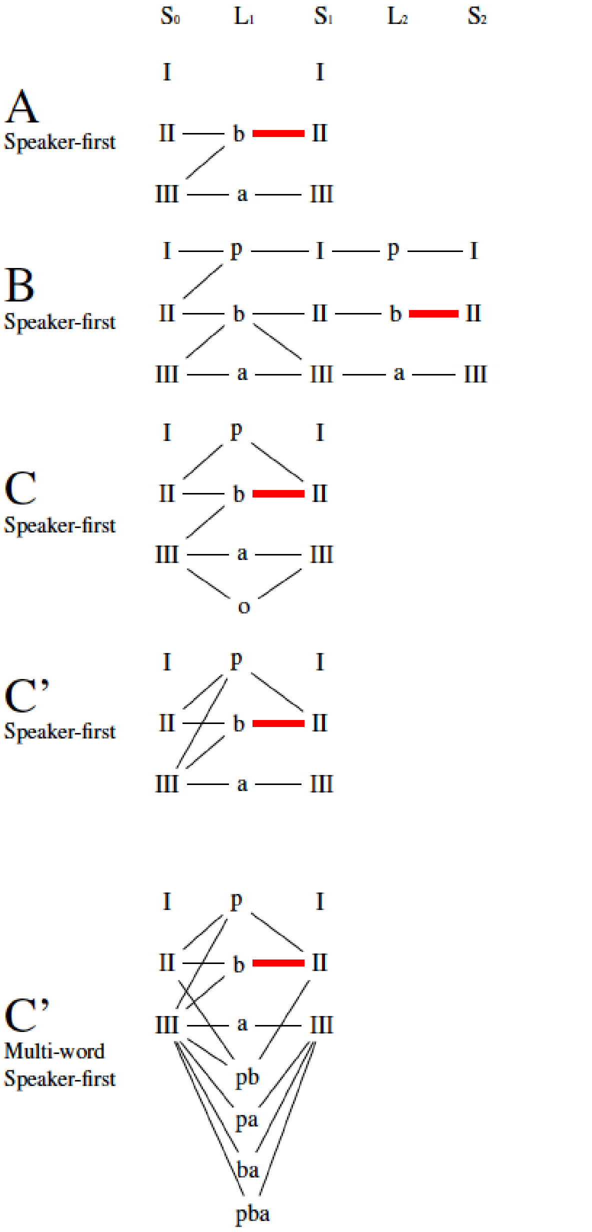


FIGURE 8: Scenarios A-C’ in IBR (speaker-first)

In the second step, a listener L_1 maps messages to states. We draw a connection from message *m* to every state *t* if in the previous step (S_0) *t* is connected to *m* and among all states connected to *m*, *t* has the lowest number of outgoing connections. In Scenario B, the message ‘apple’ is mapped to state III because in the previous iteration ‘apple’ was connected to state III and to no other state. The message ‘pear’ was connected with both state I and state II in S_0; since state I had less outgoing connections (only one connection) than state II (two outgoing connections), ‘pear’ is mapped to state I. The message ‘banana’ was connected with both state II and state III in the previous iteration; since both have an identical number of outgoing connections in the previous iteration, two each, ‘banana’ is mapped to both state II and state III in L_1 as well.

The next step is that of a speaker S_1 that maps states to messages in a way similar to L_1. We draw a connection from state *t* to every message *m* if in the previous step (L_1) *m* is connected to *t* and among all messages connected to *t*, *m* has the lowest number of outgoing connections. In Scenario B, S_1 maps state

I to the message ‘pear’ because in the previous iteration state I was connected to ‘pear’ and to no other message. State II is mapped to ‘banana’ for the same reason. As for state III, although it was connected both with ‘banana’ and with ‘apple’, the message ‘apple’ had a lower number of outgoing connections – it only had one outgoing connection (to state III) while ‘banana’ had two (to state II and state III). Thus, state III is mapped to ‘apple’.

The model continues to apply listener and speaker steps recursively until each message is mapped to at most one state. In Scenario B, this is achieved in L_2. In this step, each of the messages ‘pear’, ‘banana’, and ‘apple’ is mapped to exactly one state – states I, II, and III, respectively – because each of them was connected with only one state in S_1.

In all scenarios, IBR predicts that the message ‘banana’ would identify state II. This behavior is approximated by PSI.

This kind of reasoning can also, in principle, start with a listener step instead of a speaker step. We show listener-first computations for all scenarios in Figure 9. Applying a listener-first IBR model to our scenarios yields identical predictions in all scenarios except Scenario C. What is different about Scenario C is that in the first step, L_0, ‘banana’ is the only message that has more than one outgoing connection. This gives all other messages an advantage over ‘banana’ in step S_1, preventing it from being mapped to any state at all (because every state that has a banana also has some other fruit in it). This is a good result, since empirically ‘pick the crate with the banana’ is not grammatical in this scenario. Otherwise, IBR does predict that ‘banana’ will identify state II in all other cases in which this kind of inference does not arise.

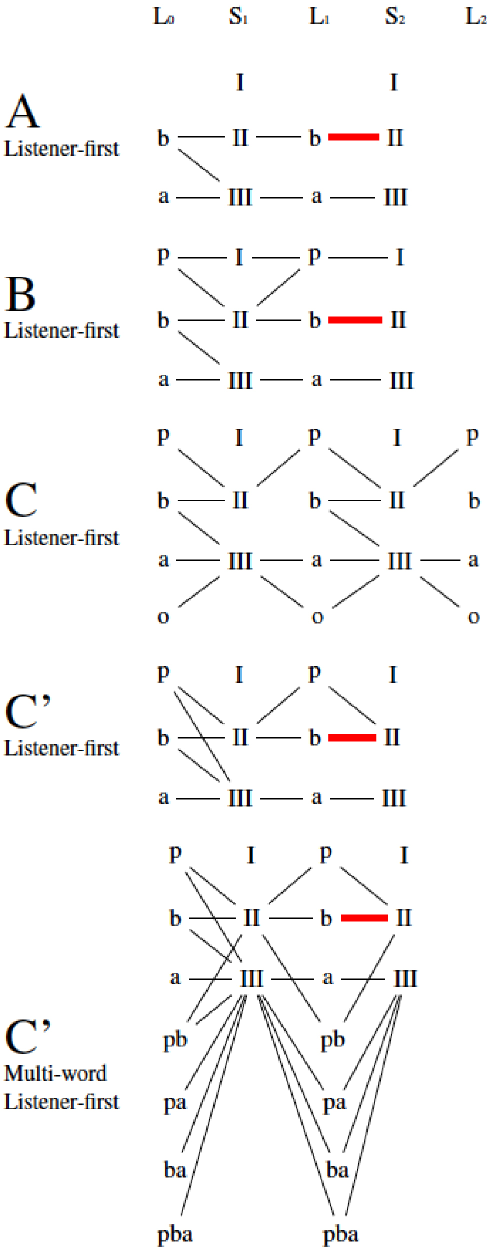


Figure 9: Scenarios A-C’ in IBR (listener-first)

3. Iterative elimination of weakly-dominated strategies

Rothschild (2013)'s IRM, based on the notion of weak dominance among strategies, iteratively eliminates strategies that a speaker and a listener could adopt based on their relative potential payoff. As we did with IBR, we will demonstrate the model's algorithm by deriving step by step its predictions for Scenario B. The derivation of Scenario B and all other scenarios is available in FIGURE 10.

The algorithm starts with a decision tree. The root node N branches into a number of nodes representing the possible states (here labeled S_I, S_{II}, and S_{III} for the three crates in each scenario). Each state node then branches into nodes corresponding to all messages that are true in that state, representing the strategies available to the speaker when trying to describe that state. Finally, each message node branches into each of the moves available to the listener when receiving the message (L_I, L_{II}, and L_{III}, one for each compatible state). Moves that correctly identify the state of the current branch have a positive payoff (we will assume a uniform payoff of 1 for each success, but see Rothschild for other possibilities).

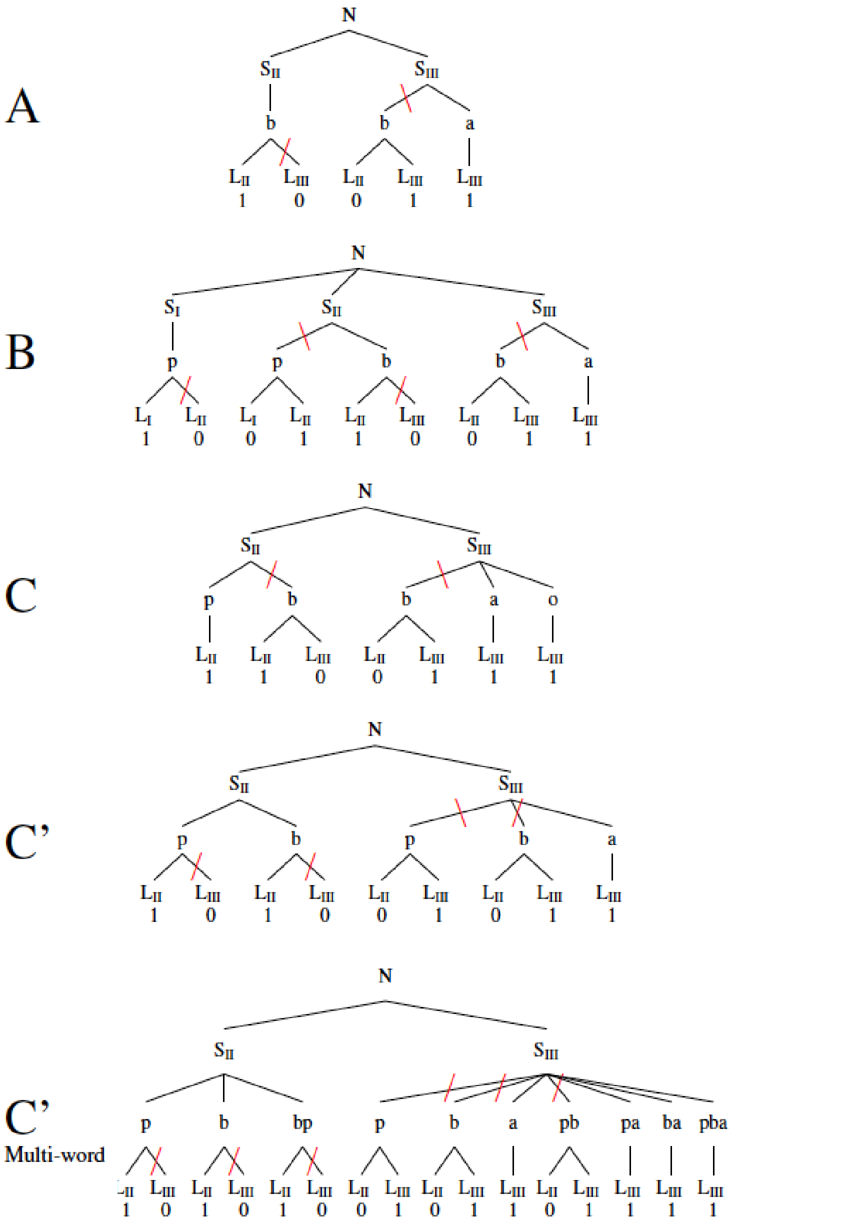


FIGURE 10: Scenarios A-C' in iterative weak-dominance reasoning

In each iteration, the players (speaker and listener) consider the potential of payoff for each strategy available to them. Elimination of strategies is determined by a relation of weak dominance among available strategies. A strategy *s* weakly dominates another strategy *s'* if *s* has at least as high a payoff as *s'* for all states of affairs (i.e. for all sequences of nodes that follow each strategy), and there is at least one state of affairs in which *s* has a strictly higher payoff than *s'*.[4]

In Scenario B the algorithm goes as follows. First, the message 'banana' yields the same *potential* payoff in both states in which it is true, S_{II} and S_{III} (payoff of 1 or 0, with the same probability). The same is true for 'pear' (in S_I and S_{II}). However, 'apple' is guaranteed to lead to a payoff of 1 in S_{III} because this is the only state that it can lead to. Since 'apple' guarantees a payoff of 1, while banana does not, and there is no other state of affairs in which 'apple' is available and has a lower payoff than 'banana', 'apple' weakly dominates 'banana' in S_{III}. Therefore, 'banana' is eliminated as a strategy in S_{III} (first elimination). In the second iteration, 'banana' is guaranteed to lead to a payoff in S_{II}, because, given the first elimination, S_{II} is the only state that 'banana' can lead to. Therefore, an interpretation of 'banana' as leading to S_{III} is weakly dominated by an interpretation as leading to S_{II} and is thus eliminated (second elimination). Next, unlike 'banana', 'pear' is not guaranteed to have a payoff of 1 in S_{II} because it is still a possible strategy in S_I as well. Since there is also no other state of affairs in which 'banana' has a lower payoff, 'banana' weakly dominates 'pear' in S_{II} and 'pear' is accordingly eliminated as a strategy in this state (third elimination). Finally, after 'pear' is eliminated as a strategy in S_{II}, it remains a strategy only in S_I, and is therefore guaranteed to have a payoff in S_I. Consequently, an interpretation of 'pear' as leading to S_I weakly dominates an interpretation of it as leading to S_{II}, and the latter is eliminated (fourth elimination).

Iterative elimination of weakly-dominated strategies, similarly to IBR, succeeds in predicting that 'banana' will not identify any state in Scenario C. The reason behind it is the same as before – at the very first iteration 'banana' is a sub-optimal message, because each state has at least one message that identifies it uniquely. Here, the message 'banana' in S_{II} is weakly dominated by the message 'pear'. In S_{III}, 'banana' is weakly dominated by 'apple' and 'orange'. In the very similar case, C' where 'banana' is not weakly dominated by another message in the first iteration, this IRM, too, predicts an attested inference from 'banana' to S_{II}..

[4] This definition of weak dominance is adapted to the current set-up, where all players are assumed to have identical payoff in all states of affairs because their goals are aligned (Rothschild 2013: p. 564). The original definition given in Rothschild (p. 448) is the following: *A strategy s for a player i weakly dominates another strategy s' if s guarantees i at least a high a payoff as s' and for some state of affairs (i.e. some possible opponent strategy and/or state of nature) s give i strictly higher payoff then s'.*