Implying or implicating *not both* in declaratives and interrogatives[1]

Matthijs WESTERA — *Universitat Pompeu Fabra*

Abstract. Both disjunctive assertions and disjunctive questions can imply "not both", i.e., that only one of the disjuncts is true. For assertions this is known to be part of what the speaker means (e.g., an implicature), whereas for questions this is instead a presupposition. This puzzle is challenging for predominant pragmatic and grammatical accounts of exhaustivity in the literature. This paper outlines a solution based on Attentional Pragmatics combined with (other) general pragmatic principles.

Keywords: exhaustivity, implicature, presupposition, disjunction, alternative question, pragmatics, intonation.

1. Introduction

Both assertions and questions exhibit so-called *exhaustivity* effects. For instance, disjunctions of either type imply 'not both':[2]

(1) a. John was at the party, or Mary. **Implied:** not both *(meant)*
b. Was John at the party, or Mary? **Implied:** not both *(not meant)*

Ideally, a theory of exhaustivity would deliver these implications in more or less the same way. At the same time, however, there is a difference: the 'not both' implication is commonly claimed to be part of what the speaker *means* in the case of assertions, like (1a), e.g., it is often called an implicature, but not in the case of questions, like (1b), where it has been claimed to be *presupposed* instead. I will review some of the evidence for this difference in section 3.

Let me briefly clarify the terminology in which the puzzle is framed. Utterances imply many things that are not part of what is meant. For instance, hearing me utter "It's raining" implies that I know those English words, hearing me say it with the accent I have will tell you I am likely Dutch. It will also imply (through an assumption of cooperativity) that I believe that it is indeed raining. Yet none of these implications are part of what I meant by the utterance. What I meant, after all, is that it is raining, not that I believe it, let alone that I know the words or am not a native speaker. This shows that explaining the presence of an implication, and explaining why it is or is not part of what is meant, are two separate things.

The challenge posed by (1) is, in a nutshell, to explain the exhaustivity implications of (1a) and (1b) in a more or less uniform way, while also accounting for the difference in status of this implication as part or not part of what is meant. The approach taken in this paper is based on two core assumptions. First, exhaustivity implications do not depend on reasoning about the informational strength of an utterance (as in standard pragmatic and grammatical approaches to exhaustivity), but on reasoning about what an utterance draws *attention* to (Westera, 2017b). Since this is a semantic/pragmatic dimension that questions and assertions share, this enables a uniform treatment of the exhaustivity implications of (1a) and (1b). Second, exhaustivity is part

[1]I would like to thank Jeroen Groenendijk and Floris Roelofsen for very useful commentary on this work in an early stage.

[2]The intended intonation here has a final low boundary tone and focus accents on each disjunct.

 In: M. Franke et al. (eds.)
Proceedings of Sinn und Bedeutung 24, vol. 2, pp.423–438. Osnabrück University.

of what is meant in (1a) but not in (1b), because it is only *relevant* in the former, and whatever a speaker means should be relevant. Of course the latter relies on several assumptions about relevance and questions vs. assertions which require substantial motivation and discussion.

Section 2 gives a summary of related work, section 3 considers some evidence from the literature corroborating the empirical puzzle, section 4 presents my account, and section 5 reflects on the required assumptions in more detail. Section 6 concludes.

2. Related work

Exhaustivity effects are often indiscriminately called 'implicatures' in the literature. Though strictly speaking signifying an (indirectly communicated) component of what is meant (Grice, 1975), the label 'implicature' is often used as synonymous with 'implication' (and 'inference'); in the grammatical approach to exhaustivity (see below) it is used even for what are treated as semantic entailments. Definitions of implicature in the literature often deviate from Grice's, and can differ even between authors whose informal conceptions of implicature seem to align. For instance, Gazdar (1979: p.38) defines it as something which is implied by the utterance of a sentence but that is not an entailment of the sentence's semantic meaning; Levinson (1983) defines it as a type of inference drawn by an audience; Gamut (1991: vol.1, p.207) define it as any logical consequence of "the conditions under which a sentence can correctly be used". This variation in definitions of implicature may be symptomatic of a number of misconceptions (see Bach, 2006). In the present paper 'implicature' is used strictly in the Gricean (1975) sense, as a certain type of component of what the speaker means.

It is not clear which of the predominant accounts of exhaustivity in the literature would be best equipped to handle the main puzzle of this paper, i.e., (1a,b). The traditional pragmatic approach is based on the Gricean maxim of Quantity: The speaker is implied to not believe that 'both' is true, because otherwise they would have asserted that instead of the (less informative) disjunction (for a critical discussion see Geurts, 2011). This is supposed to deliver the exhaustivity implications of assertions, but it does not extend to questions: In (1b) the speaker did not assert the conjunction, but they did not assert the disjunction either, and yet the latter does not end up being excluded as exhaustivity. More conceptually, only assertions involve the communication of a piece of information on which the maxim of Quantity can operate; the Gricean maxims simply do not apply to questions (requiring, instead, some other set of maxims). Moreover, pragmatic accounts have tended to gloss over the issue entirely of why exhaustivity should be part of what the speaker means, i.e., why it should be an implicature as opposed to merely a pragmatic implication: Authors simply stop as soon as the implication is delivered.

The more recent grammatical approach (e.g., Chierchia et al., 2012), which delivers exhaustivity through the insertion of grammatical operators, likewise faces a challenge. Grammatical operators make exhaustivity part of the core meaning, essentially as a semantic entailment. While this arguably predicts exhaustivity to be part of what is meant in the case of assertions, where semantic entailments end up being part of the assertion, it does not seem to apply to questions. After all, entailments do not generally 'survive' questioning force, e.g., "Does John sleep?" does not entail that John sleeps, or even presuppose it. Accordingly, it is not clear how the grammatical approach, by incorporating exhaustivity as an entailment into the semantics of a sentence, could deliver the exhaustivity implications of questions.

The explanation that I will propose is based on a different approach to exhaustivity altogether, Attentional Pragmatics (Westera, 2017b), which was motivated by a number of independent shortcomings for both aforementioned approaches – it was not aimed specifically at the puzzle at hand. The present paper is a more focused, less formal, more self-contained part of chapter 12 in Westera (2017b). For a more self-contained introduction to Attentional Pragmatics see Westera (2017a); Westera (2020a)[under review]; for a more detailed comparison in particular to the grammatical approach (focused on so-called Hurford disjunctions) see Westera (2020b)[under review].

3. The empirical puzzle

The central puzzle in this paper is that whereas (1a) and (1b) share the same exhaustivity implication, the exhaustivity is a component of what the speaker means only in (1a), not in (1b):

(1) a. John was at the party, or Mary. **Implied:** not both *(meant)*
b. Was John at the party, or Mary? **Implied:** not both *(not meant)*

The exhaustivity effects of assertions, and especially the 'not both' effect of disjunctions, have long been considered a prime example of conversational implicature, hence part of what is meant. Now, within the grammatical approach to exhaustivity this view appears to have been abandoned, but only, it seems, in favor of an account in which exhaustivity is made an even more central part of what is meant, by having it contribute to the core compositional semantics. By contrast to assertions, the exhaustivity effects of disjunctive questions have long been noted not to be part of what the speaker means but rather to be presupposed (e.g., Bartels, 1999; Aloni and Égré, 2010; Rawlins, 2008; Biezma, 2009; Biezma and Rawlins, 2012) or, a terminological variation, imposed on the common ground (Pruitt and Roelofsen, 2011).

The consensus in the literature about the status of exhaustivity (as part or not part of what is meant) has led to this being subjected to only little empirical investigation. However, a telling experiment concerning assertions is presented by Destruel et al. (2015), who use the appropriateness of "no" vs. "yes, but..." as a diagnostic for at-issueness, in examples like the following (though the experiment was in German):

(2) A: The soup is warm. (L%) **Implied:** not hot *(meant)*
B: No, it is hot. / ?? Yes, but/and it is hot.

Destruel et al. (2015) find that exhaustivity is, with overwhelming preference, contradicted by means of "no" as opposed to "yes, but..." (also "yes, and..."). From a comparison with other types of constructions they conclude, moreover, that content which is preferably denied by "no" is *at-issue content* in the sense of Simons et al. (2010), who define it (p.323) as part of what the speaker means (albeit in terms of a speaker's intentionally recognizable communicative intention; see also Goodhue and Wagner, 2018 for the relation between "yes"/"no" and at-issueness). Although Destruel et al. do not consider the 'not both' effect of disjunctions, it clearly obeys the same pattern they find for other exhaustivity effects:

(3) A: John was at the party, or Mary. (L%) **Implied:** not both *(meant)*
B: No, both were there. / ?? Yes, and/but both were there.

As for exhaustivity on questions, consider the interrogative counterpart of (3), adapted from Roelofsen and Farkas (2015):

(4) A: Was John at the party, or Mary? (L%) **Implied:** not both *(not meant)*
B: ?? Yes, not both. / ?? No, both were there. / Actually, both were there.

As observed by Roelofsen and Farkas, in this case the 'not both' implication cannot be targeted by either 'yes' or 'no'; negating it requires a marker like 'actually'. Accordingly, the 'not both' exhaustivity implication of disjunctive questions appears not to be at-issue. In line with existing characterizations (cited above), this would be because the exhaustivity is not part of what is meant but merely implied (and implied to be already known, i.e., presupposed).

Alternatively, one might try to explain the infelicity of "yes" and "no" in (4) by appealing to the fact that it is not what some would call a "yes/no question" (but cf. Bolinger, 1978; Bäuerle, 1979; Biezma and Rawlins, 2012). Note however that "yes" and "no" in (4) are meant to address not the question itself but its would-be implicature, "not both". Note, furthermore, that it is possible in principle to target the implicatures of (non-"yes/no") questions by "yes" and "no", e.g., the following dialogues are fine:[3]

(5) A: Speaking of Trump, how is he still our president?!
Implied: (e.g.,) he is a bad president.
B: Yes, he's a total failure. / No, I think he has done good things.

(6) A: What are you doing here?
B: How is that any of your business? **Implied:** it is not any of your business.
A: No, it is, I work here.

What (4) seems to show, therefore, is not the supposed infelicity of "yes" and "no" in response to supposed non-"yes/no"-questions in general – something which would of course be in need of explanation in its own right – but the absence of a suitable implicature for these response particles to target.

Now, the evidence in (4) is less strong than that in (3), because whereas being at-issue implies being part of what is meant, the opposite does not hold. For instance, parenthetical remarks clearly express something that is part of what the speaker meant but which is not the main point of the utterance. Let me employ another diagnostic to further support the assumed presupposition status of 'not both' in (4), namely the "hey wait a minute" test:

(7) A: Was John at the party, or Mary? **Implied:** not both *(not meant)*
B: Hey wait a minute, is not it possible that both were there?

[3] An anonymous reviewer notes that these are rhetorical questions, and that the assertoric nature of rhetorical questions may be to blame for the felicity of "yes"/"no" in these cases. That may be so, but in my view their assertoric nature is explained precisely by the presence of a conversational implicature (i.e., an indirectly conveyed intention to inform), not the other way around. It is difficult to find or construct a natural question that conversationally implicates something without this implicature overshadowing the main questioning act, i.e., without giving it a rhetorical flavor – but perhaps example (9) further below is such a case.

These particular examples were not intended to support any new empirical claim, but merely to illustrate the pattern of interest. The different status of exhaustivity in assertions vs. questions is widely assumed, across many different theoretical strands in the literature. In what follows I will therefore take this pattern for granted and seek an explanation.

4. The account

I will list all necessary assumptions up front, and explain how they interact to solve the central puzzle, before motivating each assumption in more detail in section 5. The assumptions are:

A. **QUDs**: The set of all 'in principle relevant' propositions, say, all pieces of information deemed worth making common ground at some point in the current discourse, is subdivided into QUDs (questions under discussion), roughly, 'ways of being relevant', of which one or several may be *active*, i.e., to be addressed by the current utterance (e.g., Roberts, 2012).

B. **I(nformation)-maxims:** A cooperative speaker will intend to communicate (or *mean*) all and only information they believe to be true that is relevant to some active QUD (essentially Grice, 1975; more precisely the definition in Westera, 2017b).

C. **A(ttention)-maxims:** A cooperative speaker will intend to draw attention to all propositions the speaker considers possible and that are relevant to some active QUD (Westera, 2017b).

D. **Symmetry:** If an active QUD contains a proposition p, then its negation $\neg p$ is also contained in an active QUD (e.g., Chierchia et al., 2012; though typically not for the same reasons, and not in the same QUD, e.g., Horn, 1989; Westera, 2017c).

E. **Closure:** QUDs are by default assumed to be closed under conjunction (e.g., Schulz and Van Rooij, 2006; Spector, 2007); this can be overruled by context and by other principles such as assumption G. below.

F. **Table:** Interrogatives normally serve to introduce a new QUD to the table, while declaratives presuppose (accommodatably) a pre-existing QUD (*ibid.*; Farkas and Bruce, 2010).

G. **Possibility:** A speaker who introduces a *new* QUD to the table should consider all propositions it contains possible (e.g., Roberts, 2012).

H. **Disjunction:** 'Contrastive' focus intonation on the disjuncts, as intended in (1a,b), indicates that the QUD is supposed to contain the individual disjuncts (e.g., Biezma and Rawlins, 2012).

I. **Low boundary tones (L%):** The L%, as in both (1a,b), indicates that the utterance is intended to comply with all the conversational maxims as far as the main QUD goes (Westera, 2018).

None of these are new to the present paper, all of them are fairly general, and all of them should sound quite plausible on the surface (though this may be subjective).

The way in which these assumptions interact to predict and explain the pattern in (1a,b) is as follows. Starting with the exhaustivity in (1a), where it is part of what is meant:

1. Given the final fall (L%), the speaker must believe that their utterance complies with the I-maxims and A-maxims relative to the main QUD (assumption I.);
2. Given the contrastive intonation, the QUD is supposed to contain the individual disjuncts (by assumption H.), and given 1. this must indeed be the case;
3. Hence their conjunction 'both' too is contained in that QUD, by the default assumption of closure of QUDs under intersection (assumption E.);
4. Compliance with the A-maxims (as per 1.) implies that the speaker must have mentioned (i.e., drawn attention to) all relevant propositions they consider possible (assumption C.);
5. Since the speaker did not mention the 'both' proposition (which they could have done by adding "or both"), they must not consider it possible; put differently, the speaker must believe 'not both', i.e., exhaustivity;
6. Because 'both' is relevant (step 2.), their negation 'not both' must be relevant too (assumption D.);
7. And therefore in (1a) the exhaustivity must be part of what is meant (B. & I.).

In a nutshell, because 'both' is relevant to the main QUD, the low boundary tone entails its exclusion, i.e., 'not both'; and since the latter is relevant too, it must be part of what the speaker means.

Things are a bit different for (1b), although the first two steps are the same:

1. Given the final fall (L%), the speaker must believe that their utterance complies with the I-maxims and A-maxims relative to the main QUD (assumption I.).
2. Given the contrastive intonation, the QUD is supposed to contain the individual disjuncts (by assumption H.), and given 1. this must indeed be the case.
3. But this time their conjunction 'both' cannot be (despite G.), because:
   i. Suppose (to obtain a contradiction) that the conjunction 'both' *is* relevant.
   ii. Since (1b) is an interrogative, it serves to introduce a new QUD (F.).
   iii. To introduce a new QUD that contains the proposition 'both', this proposition would have to be considered possible (G.).
   iv. Given 1. and the definition of the A-maxims, because 'both' was not mentioned, we know that the speaker must not consider it possible (C. &).

   There is a contradiction between iii. and iv., so the conjunction 'both' cannot be relevant.
4. The reason why 'both' is not relevant (despite D.), must be that it was not considered possible (as required by G.).
5. Put differently, the speaker must believe 'not both', i.e., exhaustivity.
6. But since 'both' is not relevant, its negation 'not both' cannot be relevant either (F.).
7. Hence in (1b) the exhaustivity cannot be part of what is meant (B. & I.).

In a nutshell, this time 'both' cannot be relevant, because if it had been relevant then the speaker – now responsible for introducing the QUD – should have included it in the QUD and hence drawn attention to it.

Even if one finds all the assumptions plausible and the logical derivation steps valid, it may be difficult to grasp the full explanation. Here are two partial paraphrases that should enable the well-rested reader to represent reasonable portions of the account in mind at once:

- Exhaustivity in assertions such as (1a) is the exclusion of relevant alternatives that were not mentioned, whereas in questions such as (1b) it is the exclusion of *ir*relevant alternatives that *would have been* relevant (hence mentioned) had they been considered possible.
- Exhaustivity is part of what is meant in (1a) but not (1b), because in the former nothing prevents the default closure of relevance under conjunction and then negation that makes exhaustivity relevant, whereas in the latter such closure is prevented by the fact that interrogatives introduce a new QUD, whose propositions must be considered possible.

Another way of gaining further insight into how the explanations work is to try to break certain steps of the derivation and see how the resulting predictions are different. For instance, recall that the explanations rely on assumptions about relevance or irrelevance of "both", namely, closure under intersection by default, which is prevented in the case of questions that introduce a new QUD. Let us try to override these assumptions by context and see what happens:

(8) A: Was John there, or Mary, or Bill?
B: John was there, or Mary.
Prediction: 'not both' not part of what B meant.

(9) A: Was John there, or Mary, or both?
B: Was John there, or Mary?
Prediction: 'not both' is part of what B meant.

In (8) A's question already implies that no combination of two or three individuals was there, in particular not both John and Mary. Accordingly, this is no longer relevant by the time B utters the disjunctive assertion, making this a case where closure under intersection does not hold even for an assertion. By contrast, (9) is a situation where A's initial question explicitly makes 'both' relevant, and speaker B arguably is not introducing a new QUD even though B is using an interrogative, making this case where the QUD for B is closed under intersection despite her uttering an interrogative. The prediction for (8), therefore, is that 'not both' is not part of what B meant, and vice versa for (9), contrary to the usual pattern.[4] This prediction seems to me immediately plausible for (8), which is a fairly straightforward interaction; but (9) is a bit odd for independent reasons – replying to a question by asking another, narrower question is non-standard – so let me clarify the intended reading of (9). In particular, the intended reading of (9) is *not* one where B asks a sub-question as part of a discourse strategy for resolving A's original question (Roberts, 2012). Rather, the intended reading is one where B is actually resolving part of A's question by conversationally implicating 'not both'. Combined

[4]In addition, it is predicted that (8) cannot imply 'not both' in the usual way – 'both' is not relevant, after all. To the extent that B's utterance does still imply 'not both', this seems to me plausibly a consequence of B's lack of protest to this presupposition of A's utterance.

with B asking the remainder of the question, this makes it feel as if B is correcting A, as if B is saying: "The 'both' option is false, so *this* is the question you should have asked". Under this reading I find the prediction that 'not both' is part of what is meant plausible.

Because all assumptions of the account are more general than the puzzle at hand, they yield predictions also for other types of utterances. For instance, I refer to Westera (2017b) for an account of simple (non-disjunctive) propositional questions as well as questions with a final high boundary tone; see Westera (2018) for an account of rising declaratives that relies on many of the same assumptions; see Meertens (2019) for an application to alternative questions.

5. A closer look at the assumptions

5.1. Assumptions B. & C.: Two sets of maxims

The explanation relies on two sets of maxims:

B. **I(nformation)-maxims:** A cooperative speaker will intend to communicate (or *mean*) all and only information they believe to be true that is relevant to some active QUD (essentially Grice, 1975; more precisely the definition in Westera, 2017b).

C. **A(ttention)-maxims:** A cooperative speaker will intend to draw attention to all propositions the speaker considers possible and that are relevant to some active QUD (Westera, 2017b).

The I(nformation)-maxims are essentially the Gricean (1975) maxims, of which I adopt the following definition (for formalization see Westera, 2017b): For each active QUD:

- **I-Quality:** Intend to communicate only propositions you believe are true;
- **I-Relation:** Intend to communicate only propositions contained in the QUD;
- **I-Quantity:** Intend to communicate the strongest propositions permitted by I-Quality and I-Relation.

The account relies on these maxims to derive that, given that 'not both' is implied to be believed by the speaker, it is part of what the speaker means (intends to communicate) if and only if it is relevant. More generally, the I-maxims predict that the contents of any implied belief will be part of what is meant if they are relevant, and not if they are not.

The A(ttention)-maxims are a more recent development (Westera, 2017a; building on Gazdar, 1979; Schulz and Van Rooij, 2006; Ciardelli et al., 2009). The main idea is that, besides intending to provide information, speakers also intentionally draw each other's *attention* to certain possibilities. As a type of communicative intention, like information-sharing, attention-drawing must be governed by a set of conversational maxims:

- **A-Quality:** Intend to draw attention only to propositions you consider possible;
- **A-Relation:** Intend to draw attention only to propositions contained in the QUD;
- **A-Quantity:** Intend to draw attention to the maximal set of propositions permitted by A-Quality and A-Relation.

For a formal definition and motivation I refer to Westera (2017b).[5]

[5]The above presentation of the A-maxims is slightly simplified compared to Westera (2017b), but not in a way

The present account relies on the A-maxims, in particular A-Quantity, for delivering (the first step towards) exhaustivity, instead of the more common pragmatic approach which relies on I-Quantity (e.g., Geurts, 2011). In Westera (2017b) this shift from I-maxims to A-maxims is argued to solve a large number of problems (see also Westera, 2020a[under review]; Westera, 2020b[under review]). For present purposes deriving exhaustivity through the A-maxims is crucial because, whereas the I-maxims apply only to assertions (i.e., there needs to be an intention to inform), the A-maxims apply to assertions and questions alike: Although questions do not (primarily) serve to provide information, they do serve to draw attention to things.

This common starting point for exhaustivity on questions and assertions notwithstanding, recall that it nevertheless derives exhaustivity for (1a) and (1b) in slightly different ways. For (1a), given that the 'both' proposition is relevant, the maxim of A-Quantity directly lets us conclude that the speaker must not consider 'both' possible. By contrast, in (1b) the maxim of A-Quantity lets us conclude only that *if* the 'both' proposition is relevant, the speaker must not consider it possible – and we relied on other assumptions and a proof by contradiction to conclude from this that the 'both' proposition in (1b) cannot be relevant.

5.2. Assumptions A., D. & E.: Questions Under Discussion (QUDs)

In any discourse, a large number of pieces of information may be considered broadly 'relevant', in the sense of being considered worth making common ground. Assumption A., repeated here, states that we tend to subdivide this potentially quite large set of pieces of information into more focused subsets, or Questions Under Discussion (QUDs; e.g., Carlson, 1983; Roberts, 2012; Ginzburg, 1996), and that each utterance is aimed only at one or a small number of QUDs.

A. **QUDs**: The set of all 'in principle relevant' propositions, say, all pieces of information deemed worth making common ground at some point in the current discourse, is subdivided into QUDs (questions under discussion), roughly, 'ways of being relevant', of which one or several may be *active*, i.e., to be addressed by the current utterance.

QUDs are not actually 'questions', at least not in the sense of either interrogative sentences, their semantic contents, or the questioning speech acts expressed by them. Rather, a QUD represents a set of pieces of information that are considered worth making common ground and which share a certain subject matter or discursive function, by virtue of which it is reasonable for a speaker to pursue them jointly with a single utterance.

Recall that the I-maxims and A-maxims above were defined relative to the utterance's QUD: It is only given a QUD that the maxims determine what is reasonable for a speaker to assert and draw attention to. This means that a different set of principles must be defined to govern which QUDs are reasonable for a speaker to pursue to begin with. Assumptions D. and E. below will be of this type. However, we do not need a full account of the principles governing QUDs for a pragmatic theory based on this notion to make predictions: An utterance itself also constrains what its QUD may be, by virtue of the maxims, but also by virtue of markers such as prosody.

Assumption D. claims that relevance is 'symmetrical':

that matters for present purposes. The full-fledged A-maxims include a requirement that every proposition in the attentional intent should be considered not only possible (A-Quality), but possible independently of any stronger proposition in the attentional intent.

D. **Symmetry:** if an active QUD contains a proposition p, then its negation $\neg p$ is also contained in an active QUD (e.g., Chierchia et al., 2012; though typically not for the same reasons, and not in the same QUD, e.g., Horn, 1989; Westera, 2017c).

The resulting account relies on this symmetry assumption to relate the (ir)relevance of 'both' to the (ir)relevance of 'not both' and, from there, to 'not both' being part or not part of what is meant. In this way Assumption D. effectively reduces the problem of why 'not both' is or is not part of what is meant, to the problem of why it is or is not relevant.

Now, assuming the symmetry of relevance can lead to the so-called *symmetry problem* for accounts of exhaustivity (e.g., Kroch, 1972; Chierchia et al., 2012; Westera, 2017c): If both a proposition and its negation are relevant alternatives for an utterance, then the same reasoning that leads to the exclusion of one (exhaustivity) will lead to the exclusion of the other – yet excluding both leads to a contradiction. The solution to this resides in realizing that, even though a proposition and its negation are both broadly speaking worth establishing, this does not mean that they will be grouped in the same QUD (Westera, 2017c) – and since the maxims operate only given a QUD, it is only the set of alternatives in the same (asymmetric) QUD that matters for exhaustivity. It is for this reason that Assumption D. crucially does not claim that a proposition and its negation are necessarily worth making common ground for the same reason, and that they would be necessarily grouped together in the same single QUD. Thus, we can adopt Assumption D. and yet maintain that the main QUD in an example such as (1a) contains only the positive propositions, i.e., that it is the QUD of who was at the party, not who was not.

More precisely, in the case of exhaustivity such as 'not both' in (1a), the assertion addresses the main QUD of who was at the party, while the 'not both' implicature addresses the secondary QUD of who was not at the party. There are several reasons for adopting this multi-QUD analysis of (1a), besides the necessity of breaking the symmetry for an account of exhaustivity, i.e., for avoiding the symmetry problem. One is that, if an utterance expresses multiple intents, such as a primary assertion and a conversational implicature, then each intent simply *must* relate to its own distinct QUD – this is because no two different intents can simultaneously comply with the maxims (as defined above) relative to the same single QUD. Another is the lack of a prosodic focus accent on the verb in (1a), which one would expect if propositions in the QUD had, besides varying in the individual, also varied in the predicate (being present vs. being absent). A third reason is that splitting the set of all broadly relevant propositions into a negative and a positive QUD is a perfectly ordinary type of discourse strategy (Roberts, 2012), and one which has certain important advantages in its own right (Westera, 2017c), primarily that it prevents the symmetry problem from arising, thus enabling communicating a large part of the answer implicitly via exhaustivity implicature.

It is also important to note that Assumption D. is perfectly compatible with the observation in Leech (1983) and Horn (1989) that our primary interests are in fact generally asymmetrical, i.e., that we tend to be interested primarily in what the world is like, not in what the world is not like. This is because even if negative information is not relevant for its own sake, it will still be relevant for secondary reasons. To see how, suppose we are primarily interested in establishing the truth of a proposition p but not its falsity, i.e., establishing the truth of a proposition but not its falsity directly helps us achieve our extra-linguistic goals. In this case, even the falsity of p is still worth establishing, because it would inform us that the discourse goal of establishing

p can no longer be achieved, and prompt us to find alternative routes to achieving our extra-linguistic goals. That is, the negations of relevant propositions are generally themselves worth establishing, not for their own sake, but because they help us prune the set of achievable goals and change our strategy accordingly. As Horn (1989) observes, the reason for sharing negative information tends to consist in the earlier consideration of its positive counterpart. Since according to this view positive and negative information tend to be relevant for very different reasons, it is natural to assume that they tend to end up in different QUDs. The fact that in the case of exhaustivity it is often the positive QUD that is explicitly addressed, aligns with the fact that we use the most prominent, most explicit intent of our utterance to address the main QUD, and less prominent intents such as implicatures to address a secondary QUD (cf. Westera, 2019).[6]

The foregoing motivation for Assumption D., in terms of pruning unachievable goals, gives us access to a more intuitive explanation of the contrast in (1a,b), i.e., a more intuitive back-story to what the above, more formal argumentation already shows. Whenever p is worth sharing, $\neg p$ is also worth sharing, not necessarily for its own sake but at the very least to prune the prior QUD in order to keep the conversational goals tidy. Since (1a) is a declarative, it serves to address a prior QUD, hence its exhaustivity implication may serve to prune it, thus be relevant and hence be intentionally communicated. By contrast, (1b) introduces a new QUD, hence there is no prior QUD in need of pruning, and the exhaustivity cannot be part of what is meant.

Assumption D. is a general assumption about the structure of relevance, not about the QUDs that subdivide it. By contrast, Assumption E. is a general assumption about QUDs:

E. **Closure**: QUDs are by default assumed to be closed under conjunction (e.g., Schulz and Van Rooij, 2006; Spector, 2007); this can be overruled by context and by other principles such as assumption G. below.

The proposed account relies on this assumption to derive that, in the case of a disjunction whose disjuncts are each relevant, 'both' is relevant too, unless there is a reason why it is not. In the case of assertions such as (1a) there is typically no such reason (though see (8)), hence 'both' is assumed to be relevant, leading to exhaustivity as the exclusion of a relevant alternative to which no attention was drawn. By contrast, in the case of questions such as (1a) there typically is such a reason (though see (9)), in which case Assumption E. serves merely to imply the existence of that reason, e.g., that the speaker introducing the QUD did not consider 'both' possible.

Default closure of QUDs under conjunction follows from a number of factors combined. First, if establishing the proposition p is a discourse goal, and establishing the proposition q is another discourse goal, then establishing their conjunction is a good way to achieve those goals (this shows why closure under *dis*junction would not be as plausible: The *dis*junction of two relevant propositions may not be informative enough to be of any use). Second, if for each of two (logically independent) propositions the goal of making it common ground is considered achievable, then by default the same is assumed for making their conjunction common ground – this is the part of Assumption E. that fails in (1b), and the reason why 'both' is not relevant in that case. Third, if two propositions each share a certain subject matter or discursive function by virtue of which they end up in the same QUD, then so does their conjunction.

[6]Well-known examples of implicature in the literature often reverse this – such as implicating "do not hire this person" by asserting something obviously irrelevant about handwriting – which is why these are often a bit funny.

An apparent counterexample to the first assumption is a case where the two goals are not independent, e.g., where establishing p is only a goal as long as q has not been established yet and vice versa. This would be an instance of the phenomenon know as "mention some" contexts, where giving *some* answer is sufficient. It is only an apparent counterexample, because it is a case where establishing p plain and simple is in fact not a discourse goal, and neither is establishing q. Rather, there is a single goal: to establish either one of p and q, say, the one that is easiest to establish (or even an arbitrary one). Accordingly, contrary to appearance the QUD should be represented not as a set containing both p and q, but as a singleton set containing an (underspecified) proposition, namely the proposition that among p and q is easiest to establish. The formalism used in Westera (2017b) to represent QUDs can handle such cases.

Assumption E. plays a role in case of disjunctions and their 'not both' implication, where it crucially relates the relevance of the disjuncts to that of their conjunction 'both'. For the account to generalize to other kinds of exhaustivity implications Assumption E. would have to be generalized. For instance, for the account to handle cases where "some" implies "not all", one would need to add the assumption that, by default, if the proposition expressed by means of "some" is relevant then so is the proposition expressed by means of "all" – this is essentially the common assumption that "some" and "all" form a 'scale' (Horn, 1972; Geurts, 2011). The robustness of this type of assumption to exceptions will be different for the various possible triggers of exhaustivity (e.g., there are uses of "some" where replacing it by "all" would not make for a relevant contribution at all), and the robustness of predictions of the account will vary accordingly.

5.3. Assumptions F. & G.: Asking questions

The account relies on two fairly minimal assumptions about questions, the first about the difference between questions and assertions that is ultimately responsible for the contrast in (1):

F. **Table:** Interrogatives normally serve to introduce a new QUD to the table, while declaratives presuppose (accommodatably) a pre-existing QUD (Farkas and Bruce, 2010).

As I meant to show with examples (8) and (9), Assumption F. is again a type of default assumption, permissive of exceptions with the predictions of the account changing accordingly.

Assumption F. predicts that someone who utters an interrogative is responsible for the choice of QUD in a way that someone who utters a declarative is not. Assumption G. in turn spells out one consequence of this responsibility:

G. **Possibility:** A speaker who introduces a *new* QUD to the table should consider all propositions it contains possible.

Roberts captures this in her definition of "QUD" (2012, p.14), and the first half reappears in her "Pragmatics of Questions" (p.22), but this assumption generalizes beyond the notion of QUDs in communication to goals more generally: One should not set new goals if one already knows that these are unachievable. The same idea occurs in Cohen and Levesque's (1990) formal theory of goals (their "realism" constraint, p.227).

For the present account, Assumption G. serves to effectively weaken Assumption E. of closure under conjunction, to closure under conjunction *as far as these conjunctions are considered possible*. This entails for the disjunctive question in (1b) that the reason for not including 'both'

in the QUD is that the speaker must not consider it possible, which is how the exhaustivity implication 'not both' of (1b) is derived. (By contrast, for (1a) the 'not both' implication followed more directly from the relevance of 'both' and the A-maxims.)

Strictly speaking, Assumption G. is a bit too weak (and accordingly Assumption E., which Assumption G. in turn is supposed to weaken, is a bit too strong). One should consider possible not just the propositions in a QUD themselves, but also their being made common ground, which is not the same. After all, a proposition can be considered possible even if making it common ground is considered unachievable, namely, if it is known that no one knows whether the proposition is in fact true. Strengthening Assumption G. in this way would subtly change the predictions of the account: The question in (1b) would primarily imply that the speaker considers the goal of making 'both' common ground unachievable, not necessarily that they believe that 'both' is false. In Westera (2017b) I explore certain consequences of this nuance, and show that the following assumption could be added to strengthen this prediction: that speakers tend to introduce only QUDs for which they consider it at least possible that each of their propositions will be completely resolved (affirmed or negated).

5.4. Assumptions H. & I.: Prosody

A crucial ingredient of the account is that the utterances are taken to comply with the maxims. Compliance with the I-maxims is needed to justify the assumption that the information that is both taken to be true and relevant is part of what the speaker meant. Compliance with the A-maxims is required to derive the exhaustivity implication 'not both' to begin with (in particular the maxim of A-Quantity), where the A-maxims (unlike the I-maxims) crucially apply to questions as well as assertions. Compliance with the maxims is often simply assumed as the starting point of pragmatic explanations, but this is justified only in systems where compliance with the maxims is always possible (i.e., where they are defined in such a way that they never *clash*). The current definition of the maxims does permit clashes, however, and this means that compliance with the maxims cannot simply be taken for granted.[7]

A solution to this puzzle is to assume that speakers themselves indicate the status of the maxims. Indeed, my theory of Intonational Compliance Marking (Westera, 2013, 2017b) entails that speakers of English use prosody, in particular right boundary tones H% and L%, to indicate whether they believe all the maxims are complied with relative to the main QUD. For present purposes what matters are the low boundary tones:

I. **Low boundary tones (L%):** The L%, as in both (1a,b), indicates that the utterance is intended to comply with all the conversational maxims as far as the main QUD goes.

My dissertation (Westera, 2017b) presents the Intonational Compliance Marking theory in detail. But the core idea that rises and falls indicate non-compliance and compliance with the maxims is essentially just a way of making existing characterizations in the literature more precise: characterizations of rises/falls as indicating the pragmatic incompleteness/completeness of an utterance (e.g., Bolinger, 1982; Pierrehumbert and Hirschberg, 1990; Bartels, 1999).

[7]If instead the maxims were to be redefined to avoid clashes, they would be either too restrictive and too often prevent speakers from making a contribution, or they would be too weak to be able to deliver exhaustivity.

A known feature of QUDs is that their shape is reflected, in English and many other languages, by prosodic focus (Beaver and Clark, 2009). Assumption H. captures one aspect of this:

H. **Disjunction:** 'Contrastive' focus intonation on the disjuncts, as intended in (1a,b), indicates that the QUD is supposed to contain the individual disjuncts.

This assumption (shared by, e.g., Biezma and Rawlins, 2012) can in fact be derived from a standard view on focus: Focus intonation on the two disjuncts is compatible either with (a QUD paraphrasable as) a single-wh question "Who was at the party?", or with a disjunctive multi-wh question "Who was at the party or who was at the party?". The latter is a rather strange creature; it would make sense only in certain very specific contexts, e.g., as an echo question, or with a different implicit domain restriction for each disjunct. In the absence of the contextual factors required for this more specialized interpretation, the default interpretation of a disjunction with focus on both disjuncts will therefore be the first one, which corresponds to Assumption H.

Note that Assumption H. is phrased rather cautiously: The QUD is *supposed* to contain the disjuncts. This is because it is possible for a speaker to be uncertain about what the QUD is. In the derivations in section 4 I went from their supposed relevance (Assumption H.) to their actual relevance in a single step by appealing to compliance with the maxims (Assumption I.), glossing over the following derivation. First, the speaker must have intended to draw attention either to the two disjuncts, or only to the disjunction as a whole – the form of the utterance does not fully determine this. Then, from the supposed relevance of the disjuncts we can conclude, through compliance with the maxim of A-Quantity, that the speaker must have intended to draw attention to them (even mere possible relevance is enough for A-Quantity to demand that attention be drawn to it). Lastly, from the latter, combined with compliance with the maxim of A-Relation (to draw attention only to things that are relevant), we can conclude that the speaker must consider the two disjuncts to be indeed relevant.

Details like the latter may seem excessive – it is tempting to just take for granted that the disjuncts are relevant, or that the disjunction is used with the intention of drawing attention to the them – but these simplifications would result in false predictions even for subtly different utterances, e.g., disjunctions with a single, 'broad' focus or disjunctions with a final rise (H%).

6. Conclusion

In a nutshell, the implication "not both" is part of what is meant for declaratives but not interrogatives, because its positive counterpart "both" is part of the QUD for declaratives but not interrogatives. This solves a long-standing puzzle, one which is challenging for predominant pragmatic and grammatical approaches to exhaustivity, through the interaction of general pragmatic principles. It highlights that to derive an implication and to explain its being part or not being part of what is meant are two separate issues, the latter of which has been unduly neglected.

References

Aloni, M. and P. Égré (2010). Alternative questions and knowledge attributions. *The Philosophical Quarterly 60*(238), 1–27.

Bach, K. (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*, pp. 21–30. John Benjamins Publishing Company.

Bartels, C. (1999). *The intonation of English statements and questions: a compositional interpretation*. Routledge.

Bäuerle, R. (1979). Questions and answers. In R. Bäuerle, U. Egli, and A. von Stechow (Eds.), *Semantics from Different Points of View*. Berlin: Springer.

Beaver, D. and B. Clark (2009). *Sense and Sensitivity: How Focus Determines Meaning*. Number 12 in Explorations in Semantics. John Wiley & Sons.

Biezma, M. (2009). Alternative vs polar questions: the cornering effect. In S. Ito and E. Cormany (Eds.), *Semantics and Linguistic Theory (SALT) 19*.

Biezma, M. and K. Rawlins (2012). Responding to alternative and polar questions. *Linguistics and Philosophy 35*(5), 361–406.

Bolinger, D. L. (1978). Yes-no questions are not alternative questions. In *Questions*, Number 1 in Synthese Language Library, pp. 87–105. Dordrecht, Holland: Reidel Publishing Company.

Bolinger, D. L. (1982). Intonation and its parts. *Language 58*, 505–533.

Carlson, L. (1983). *Dialogue Games*. Dordrecht: Reidel.

Chierchia, G., D. Fox, and B. Spector (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In C. Maienborn, P. Portner, and K. von Heusinger (Eds.), *Semantics: An International Handbook of Natural Language Meaning*, Volume 2, pp. 2297–2332. Mouton de Gruyter.

Ciardelli, I., J. Groenendijk, and F. Roelofsen (2009). Attention! *Might* in inquisitive semantics. In S. Ito and E. Cormany (Eds.), *Semantics and Linguistic Theory (SALT) 19*.

Cohen, P. R. and H. J. Levesque (1990). Intention is choice with commitment. *Artificial Intelligence 42*, 213–261.

Destruel, E., D. Velleman, E. Onea, D. Bumford, J. Xue, and D. Beaver (2015). A cross-linguistic study of the non-at-issueness of exhaustive inferences. In F. Schwarz (Ed.), *Experimental Perspectives on Presuppositions*, Number 45 in Studies in Theoretical Psycholinguistics, pp. 135–156. Springer International Publishing.

Farkas, D. and K. Bruce (2010). On reacting to assertions and polar questions. *Journal of Semantics 27*, 81–118.

Gamut, L. T. F. (1991). *Language, Logic and Meaning (vols. 1 and 2)*. Chicago University Press.

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

Geurts, B. (2011). *Quantity Implicatures*. Cambridge University Press.

Ginzburg, J. (1996). Dynamics and the semantics of dialogue. In J. Seligman and D. Westerståhl (Eds.), *Language, Logic, and Computation*, Volume 1.

Goodhue, D. and M. Wagner (2018). Intonation, yes and no. *Glossa: a journal of general linguistics 3*(1).

Grice, H. P. (1975). Logic and conversation. In P. Cole and J. Morgan (Eds.), *Syntax and Semantics*, Volume 3, pp. 41–58.

Horn, L. R. (1972). *On the Semantic Properties of Logical Operators in English*. Ph. D. thesis, University of California Los Angeles.

Horn, L. R. (1989). *A Natural History of Negation*. Chicago: University of Chicago Press.

Kroch, A. (1972). Lexical and inferred meanings for some time adverbs. *Quarterly Progress Reports of the Research Laboratory of Electronics 104*, 260–267.

Leech, G. (1983). *Principles of Pragmatics*. London: Longman.

Levinson, S. C. (1983). *Pragmatics*. Cambridge textbooks in linguistics. Cambridge University Press.

Meertens, E. (2019). How prosody disambiguates between alternative and polar questions. In *Proceedings of the 22nd Amsterdam Colloquium*.

Pierrehumbert, J. B. and J. Hirschberg (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack (Eds.), *Intensions in communication*, pp. 271–311. Cambridge, MA: MIT Press.

Pruitt, K. and F. Roelofsen (2011). Disjunctive questions: prosody, syntax, and semantics. Ms. presented at a seminar at the Georg August Universität Göttingen; retrieved from `https://illc.uva.nl/inquisitive-semantics`.

Rawlins, K. (2008). *(Un)conditionals: an investigation in the syntax and semantics of conditional structures*. Ph. D. thesis, University of California Santa Cruz.

Roberts, C. (2012). Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics 5*, 6:1–96. Publication of a 1996 manuscript.

Roelofsen, F. and D. F. Farkas (2015). Polarity particle responses as a window onto the interpretation of questions and assertions. *Language 91*(2), 359–414.

Schulz, K. and R. van Rooij (2006). Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation. *Linguistics and Philosophy 29*(2), 205–250.

Simons, M., J. Tonhauser, D. Beaver, and C. Roberts (2010). What projects and why. In B. Jackson and T. Matthews (Eds.), *Semantics and Linguistic Theory (SALT) 20*, pp. 309–327.

Spector, B. (2007). Scalar implicatures: Exhaustivity and gricean reasoning. In M. Aloni, A. Butler, and P. Dekker (Eds.), *Questions in Dynamic Semantics*, pp. 225–250. Elsevier.

Westera (2017a). An attention-based explanation for some exhaustivity operators. In *Proceedings of Sinn und Bedeutung*. University of Edinburgh.

Westera, M. (2013). 'Attention, I'm violating a maxim!' A unifying account of the final rise. In R. Fernández and A. Isard (Eds.), *Proceedings of the Seventeenth Workshop on the Semantics and Pragmatics of Dialogue (SemDial 17)*.

Westera, M. (2017b). *Exhaustivity and intonation: a unified theory*. Ph. D. thesis, submitted to ILLC, University of Amsterdam.

Westera, M. (2017c). QUDs, brevity, and the asymmetry of alternatives. In *Proceedings of the 21st Amsterdam Colloquium*. University of Amsterdam.

Westera, M. (2018). Rising declaratives of the quality-suspending kind. *Glossa: a journal of general linguistics 3(1)*.

Westera, M. (2019). Rise-fall-rise as a marker of secondary QUDs. In D. Gutzmann and K. Turgay (Eds.), *Secondary content: The semantics and pragmatics of side issues*. Leiden: Brill.

Westera, M. (2020a). Attentional pragmatics: a new pragmatic approach to exhaustivity. Under review; manuscript available from `https://semanticsarchive.net/Archive/2FlMzkzM/`.

Westera, M. (2020b). Hurford disjunctions: an in-depth comparison of the grammatical and the pragmatic approach. Under review; manuscript available from `https://semanticsarchive.net/Archive/jk5YzYyO/`.