QUDs, brevity, and the asymmetry of alternatives

Matthijs Westera

Universitat Pompeu Fabra
matthijs.westera@gmail.com

Abstract

Exhaustivity is typically explained in terms of the exclusion of unmentioned alternatives. For this to work, the set of alternatives must be *asymmetrical*, lest both a proposition and its negation get excluded, yielding a contradiction (the *Symmetry Problem*). Since exhaustivity is regularly observed, these alternative sets must tend to be asymmetrical, and this requires an explanation. Existing explanations are based on considerations of brevity, but these run into certain problems. A new solution is proposed, explaining the asymmetry of alternatives in terms of the fact that discourse strategies with asymmetrical questions under discussion (QUDs) are favored because they allow part of the answer to be communicated implicitly, namely as an exhaustivity implicature.

1 Introduction

In (1), B's answer with falling intonation can be interpreted exhaustively:

(1) A: Who (of your friends John, Mary, Bill, Sue and Chris) was at the party?
B: John and Bill were there. → not Mary, not Sue, not Chris

Exhaustivity is typically explained in terms of the exclusion of unmentioned (and non-entailed) alternatives. For this to work, the set of alternatives must be *asymmetrical*, lest both a proposition and its negation get excluded, which would yield a contradiction. For instance, in (1) the set must contain only people's presences, not people's absences, for excluding both Mary's presence and Mary's absence yields a contradiction. Since exhaustivity is regularly observed, these alternative sets must tend to be asymmetrical, and this requires an explanation. This was pointed out by Kroch [21] (and subsequently discussed in [12, 25], among others) and it is currently known as the *Symmetry Problem* (attributed to MIT course notes by Heim and von Fintel).

Gazdar [9] proposed a potential solution to the Symmetry Problem in terms of *scales*: lexical entries would be associated with certain intrinsically asymmetrical scales of alternatives [14]. However, Russell [29] points out that scales aren't really a solution to the Symmetry Problem unless one explains why scales are the way they are, and why they should be what drives exhaustivity; and Geurts [10] notes that there is only very little explicit reflection on what scales are supposed to be. One option is to conceive of scales as indirect representations of *what is typically relevant* given that a certain lexical expression is used (following [22], [10] and presumably [14]). Another is to conceive of scales as representations not of what is typically relevant but of what is *actually* relevant for a given utterance (following, I believe, [13] and [24]; in this role scales are also called "Hirschberg scales" or "ad hoc scales" [18]). But regardless, as several authors note, scales don't explain the asymmetry they describe (e.g., [15, 29, 10]).

Previous explanations for the asymmetry of alternative sets have relied on considerations of *brevity* (e.g., [26, 1, 16, 22]). To illustrate:

(2) A: Were (all of) your friends at the party?
B: *Some* of them were there. → not all

To explain the exhaustivity, the alternative set must contain "all (of them were there)", but not its mirror image "some but not all". The brevity-based approach proposes that this is because "some but not all" is too cumbersome to express, unlike "all" or "some". For (1) a similar explanation may be given by assuming that "weren't" is significantly more cumbersome to express than "were". This approach faces some challenges that will be discussed in more detail in section 3. One of these is the possibility of exhaustivity on negative answers (again depending on intonation):

(3) B: (Of your friends,) *Mary* and *Sue* weren't there. → the others were there.

For the brevity-based explanation to explain this, "weren't there" must now be *less* complex than "were there", the opposite of what was required above.[1]

This paper proposes that challenges for existing brevity-based accounts, like (3), stem from relying on the wrong kind of 'brevity'. A division of pragmatic labor exists between choosing conversational goals and selecting the means for pursuing them, and, crucially, both choices may be guided by considerations of brevity. Previous approaches have concentrated exclusively on the brevity of one utterance compared to alternative utterances that would address the same goals; this paper proposes to consider also the brevity benefits of pursuing one set of goals rather than another, or, as I will say following much work in pragmatics, the brevity benefits of pursuing one Question Under Discussion (QUD) rather than another (e.g., [27]). The explanation for the asymmetry of alternative sets is then, in a nutshell, that pursuing asymmetrical QUDs rather than their symmetrical counterparts favors brevity.

Although considerations of brevity and notions like conversational goals or QUDs are fundamentally pragmatic, and I will assume a pragmatic source of exhaustivity in what follows, the proposed explanation of the asymmetry of alternative sets is intended to apply independently of whether exhaustivity is derived pragmatically or as part of some linguistic convention. This is because the explanation can be understood either as a synchronic (speaker-level) rationalization for pursuing asymmetrical QUDs, or as a diachronic (population-level) explanation for asymmetrical lexical scales, when these are conceived of as conventionalizations of *typical* QUDs (the perspective on scales taken in, e.g., [10, 22]). Previous brevity-based explanations, and most pragmatic explanations, likewise permit both interpretations.[2]

2 The solution: splitting a symmetrical QUD

Suppose for the sake of argument that A's interests in (1) are symmetrical, and that A's question introduces a symmetrical QUD, e.g.:

$$\text{QUD} = \{Pj, Pm, Pb, Ps, Pc, \overline{Pj}, \overline{Pm}, \overline{Pb}, \overline{Ps}, \overline{Pc}\}$$

(Additional constraints like closure under union and intersection may also be assumed, but will not matter in what follows.) In fact I think that A's interrogative in (1) does not make a particularly strong case for the QUD being symmetrical. But I also think that one can add "and who wasn't there?" to A's interrogative, which is more suggestive of a symmetrical QUD, without this making an exhaustive interpretation on B's response impossible. So, for the sake of

[1] Katzir's [19] 'complexity' does achieve this, but as a consequence it cannot be understood as implementing a global, pragmatic preference for brevity – and Katzir notes it is not intended as such. See also footnote 4.

[2] The possibility that asymmetrical QUDs have conventionalized as scales doesn't tell us much about *how* exhaustivity would arise conventionally, e.g., by means of certain constraints on the use of invisible operators [7] or otherwise.

argument, let us assume a symmetrical QUD in (1). In addition, let us assume that B's response in example (1) targets the same, symmetrical QUD supposedly introduced by A's question.

In addition, I will assume that exhaustivity follows in some way from compliance with the maxims, whether through considerations of Quantity [30], or, if one pleases, through an operator or other type of linguistic convention plus (often left implicit by proponents) Manner and Quality. If indeed exhaustivity follows in some way from the conversational maxims, then the assumption that B's response in example (1) would comply with the maxims relative to the symmetrical QUD leads to a contradiction – this is the Symmetry Problem. Put differently: B's response cannot address such a QUD while complying with the maxims. Although speakers may in principle violate maxims, namely in case of a clash, as Grice [11] noted they must not do so silently, lest they be liable to mislead; and I have proposed elsewhere [31] that maxim violations are signaled prosodically, by a final rising contour (or in written text by, e.g., "..."). In the absence of such cues, as in example (1), only one conclusion is possible: contrary perhaps to appearances, B's answer must be aimed at a different QUD, i.e., different from the symmetrical QUD supposedly introduced by A.

This conclusion, that speaker B in (1) cannot be addressing the symmetrical QUD, is in a way just a restatement of the Symmetry Problem. But this restatement could also be the first part of the solution, provided we can answer the important issues this raises:

(i) Which QUD is (or which QUDs are) addressed by speaker B in (1), if not the symmetrical QUD supposedly introduced by speaker A?

(ii) Why was this a rational choice of QUD for B?

(iii) How can an addressee (e.g., speaker A) figure this out, accommodate the new QUD(s) and compute the right inferences?

To complete the explanation, then, I will try to answer each of these questions in a thorough way.

Question (i): Which QUDs are addressed by B? I propose that for some reason (see question (ii) below) speaker B in (1) decided to split the prior QUD, if it was indeed symmetrical, into two asymmetrical QUDs, which I will denote by QUD^+ and QUD^-:

$$\text{QUD}^+ = \{Pj, Pm, Pb, Ps, Pc\} \qquad \text{QUD}^- = \{\overline{Pj}, \overline{Pm}, \overline{Pb}, \overline{Ps}, \overline{Pc}\}$$

It should be uncontroversial that an utterance that addresses multiple QUDs should convey an appropriate communicative intent for each QUD, i.e., an intent which complies with the maxims relative to that QUD. This explains why B's response in (1) would be fine with the assumed QUDs, in the following way:

1. B's primary (asserted/explicit) intent is that John and Bill were at the party, which can safely comply with the maxims relative to QUD^+;
2. because QUD^+ is asymmetrical, compliance with the maxims of the primary intent relative to this QUD safely implies exhaustivity, i.e., that according to the speaker Mary, Sue and Chris were absent;
3. the exhaustivity implication in turn enables the clear communication of a secondary intent, i.e., an conversational implicature, namely that Mary, Sue and Chris were absent;[3]

[3]The distinction between implication and implicature is important [2]: what is implied is not necessarily implicated (meant), and what is implicated is not necessarily implied to be true (but, typically, implied to be held true by the speaker).

4. the secondary intent can safely comply with the maxims relative to the other asymmetircal QUD^{-}.

That is, instead of addressing "who was there and who wasn't there?", for some reason speaker B decided to address only the positive half explicitly, enabling B to address the negative half implicitly by means of an exhaustivity implicature.

Some authors may disagree with my invocation of asymmetrical QUDs, for they may conceive of the Symmetry Problem as a deeper problem, due to a speaker's interests or 'relevance' being necessarily symmetrical (e.g., [4]; [8]; [7]). There are no good arguments for this position in the literature, and it contrasts with the better-argued stance of Horn [15] and Leech [23] that speakers tend to be much more interested in what there is than in what there isn't – an instance of what Horn calls the *Asymmetry Thesis* [17]. Moreover, in life we rely on default assumptions all the time, only the negations of which will be worth asserting (because only the negations of our default assumptions have the potential to change our plans and behaviors). But all of this is arguably irrelevant, because even if a speaker's interests happen to be symmetrical now and then, that doesn't mean that the QUDs must in that case also be symmetrical. The reason is that the choice of QUDs depends not just on what information is deemed interesting/relevant, but also on what the best discourse strategy is for getting this information into the common ground in an clear, orderly and efficient way [27]. It may well be rational to organize one's occasionally symmetrical interests into asymmetrical QUDs – which brings us to the following question.

Question (ii): Why was this QUD-split rational? Splitting a QUD into two is an ordinary *discourse strategy* [27], and one which in this particular situation offers a substantial benefit over simply addressing the original, symmetrical QUD. The benefit is that addressing an asymmetrical QUD enables an exhaustivity implicature, unlike the original symmetrical QUD (that's the Symmetry Problem) and that the exhaustivity implicature enables the speaker to address half of the original QUD (namely, the other asymmetircal half) implicitly, which greatly benefits brevity. In a sense, the symmetry problem solves itself: an asymmetrical QUD is favored precisely because the symmetrical QUD prevents an exhaustivity implicature. Interestingly this explanation, unlike existing brevity-based accounts, generalizes to exhaustivity on negative answers like (3), repeated here:

(3) B: (Of your friends,) *Mary* and *Sue* weren't there. → the others were there.

There is no reason why a speaker shouldn't decide to address the negative QUD explicitly, explicating who was *absent* and implicating that the others were *present*.

Now, this explanation doesn't mean that this kind of QUD-split is always appropriate. For instance, if speaker A is ticking boxes on a checklist of individuals, it might be better for B to address the entire original QUD explicitly, and in the precise order of the checklist:

(4) B: John was there, Mary wasn't, Bill was, Sue wasn't, and Chris wasn't.

Moreover, addressing half of the original QUD implicitly may not be a good idea in cases where the domain of relevant individuals is not entirely clear, which would compromise the clear communication of the exhaustivity implicature. But in other circumstances B's decision in (1), to split the QUD, seems to be perfectly rational.

In general, the more people were present (and also the greater the domain of A's inquiry) the greater the brevity benefit of explicating only who was absent. But other factors will also play a role, such as which of the two properties, being present or being absent, is the most salient in the broader context, e.g., the negative answer in (3) seems particularly natural if B normally takes

attendance by writing down only the names of those who are absent. Other factors that may play a role are, for instance, which of the two predicates is the most readily lexically accessible, and which of the individuals' names B is more likely to mispronounce. But for present purposes such complications can be set aside, because the main brevity benefit is more general: relying on conversational implicature for conveying part of the answer benefits brevity regardless of the particular lexical entries involved, because an implicature is, indeed, implicit. In this regard my explanation is crucially different from existing brevity-based approaches to the Symmetry Problem, see section 3.

A final remark regarding this issue, before turning to question (iii). While the brevity benefit alone may explain why B chose to split the prior QUD into two halves, it does not explain why it should be split into a positive QUD and a negative QUD, i.e., QUD^+ and QUD^-, rather than, e.g.:

$$\text{QUD}_1 = \{Pj, \overline{Pm}, Pb, \overline{Ps}, Pc\} \qquad \text{QUD}_2 = \{\overline{Pj}, Pm, \overline{Pb}, Ps, \overline{Pc}\}$$

After all, this split could have offered, depending on who was actually at the party, a similar brevity benefit as the split into QUD^+ and QUD^- (though with different predicted exhaustivity implicatures). One reason why the above split may be dispreferred is that the resulting QUDs are more complex: the propositions in QUD^+ and QUD^- vary only along a single dimension, i.e., the individual, whereas the propositions in QUD_1 and QUD_2 vary along two dimensions, i.e., the individual, and whether they were absent or present – and they vary in a rather unpredictable way, including some but not all combinations of individual and absence/presence. This added complexity would compromise clarity: an addressee may not be able to figure out which of many possible asymmetrical-but-mixed QUDs the speaker may be addressing (also if we take prosodic focus into account, see below). I take this to explain why the QUD-split must be as assumed, i.e., into QUD^+ and QUD^-.

Question (iii): How could an addressee figure this out? The main answer to this question is that B's response in (1) *must* be aimed at a different QUD, because it would have violated a maxim relative to the symmetrical QUD – and addressees should recognize this. Which different QUD(s) speaker B may be addressing is constrained, in turn, by the notion of discourse strategy: it must be some combination of QUDs that together cover the original one, and we have already explained why the assumed split into a positive and a negative QUD is favored over the more arbitrary mixes.

Besides these general pragmatic considerations, in spoken language addressees may of course also rely on *prosodic focus* for identifying the QUD. For B's response in (1) to imply exhaustivity, it should have a pitch accent on the individuals' names but not on the predicates. Let us assume that accent placement reflects, through focus structure in the usual manner (e.g., [28, 3]), only the primary QUD, i.e., the QUD that is explicitly addressed. The focus structure of (1) will then help an addressee to figure out that the primary QUD of B's response is the asymmetrical, positive one. In contrast, if B had been addressing the symmetrical QUD, or a strange mixture like QUD_1 or QUD_2 above, B should have used either broad focus (i.e., on the entire sentence, which would normally entail an accent at least on the predicates) or multiple foci, i.e., both the individuals' names and the predicates. To illustrate, example (4) does address the symmetrical QUD, and indeed an intonation contour with accents on both the names and the predicates seems the most natural there.

Summing up, pragmatic accounts of exhaustivity that predict a contradiction relative to a symmetrical QUD, enable a solution to the Symmetry Problem precisely because this contradiction means that a QUD-shift must have taken place. The QUD-shift can be understood in

terms of a rational discourse strategy, namely, that of splitting a symmetrical QUD into two asymmetrical QUDs, which offers a brevity benefit by virtue of enabling part of the answer to be communicated via exhaustivity implicature.

3 A closer look at previous brevity-based approaches

Approaches based on brevity would attempt to solve the Symmetry Problem in cases like (1) by assuming that "John wasn't there" is a more complex expression than "John was there", and likewise for the other individuals. This would provide speakers with an excuse for omitting Mary's absence but not for omitting Mary's presence, thus breaking the problematic symmetry (e.g., [22]). Several authors have tried to define an appropriate notion of brevity/complexity, for instance in terms of number of syllables [26] or degree of lexicalization [1, 16].[4] But this type of approach runs into several problems.

First, although it seems true that "wasn't" is more complex than "was", what this approach crucially need to assume is that the purported difference in complexity would be *sufficiently large* for it to matter. That is, the difference should be sufficiently large to provide speakers with an excuse for not mentioning certain relevant propositions, *and* for this omission to not cause any confusion among the addressees. Moreover, this would have to hold even if in other regards the speaker appears not to care too much about brevity, e.g.:

(5) Well, that is a most interesting question indeed, and I am delighted to be able to assist. Of your dear friends, John was there, and Mary was there.

After all, this seems to imply exhaustivity in the same way. Furthermore, this approach would have to assume a similarly significant brevity difference even between "was absent" and "was present", by which "wasn't there" and "was there" can be replaced in the relevant examples without changing the exhaustivity implications.

But even (or especially) if it can be shown that "wasn't" (or "absent") is *sufficiently* more complex than "was" (or "present"), there is still the problem of exhaustivity on negative answers like (3). In order for previous brevity-based accounts to explain this, "was" would have to be *more* complex than "wasn't" – the converse of what is needed for (1). What example (3) shows is that a brevity-based solution to the symmetry problem that feeds only on intrinsic properties of particular lexical entries is inadequate; rather, there must be a contextual parameter of, say, "mentionworthiness", that has nothing necessarily to do with intrinsic brevity or complexity. A similar criticism is voiced by Matsumoto [25], based on cases where a simple expression and a more complex expression are used together, e.g.:

(6) B: It was warm today, and a little bit more than warm yesterday.

Matsumoto observes that the utterance implies that it was not a little bit more than warm today, despite this being expressible only by a more complex utterance. In response, Katzir [19] proposes that sometimes complex expressions can be used in spite of their complexity, and that one can find out whether complex expressions can be used by checking whether the

[4] Katzir [19] tries to filter something like relevance in terms of a measure of grammatical complexity, in a way that would superficially seem to belong in the same strand as the aforementioned approaches. However, Katzir does not intend this to be part of a pragmatic explanation, and indeed it is difficult to see how it could be. Katzir's measure of grammatical complexity is defined in terms of whether certain permissible substitutions enable one to transform one sentence into another. A consequence of this is that which of two expressions counts as more complex in Katzir's sense can depend on which expression was actually uttered. Although we can see the appeal of this proposal within the otherwise unappealing grammatical approach to exhaustivity, it does not follow from a global pragmatic preference for brevity.

utterance itself contains such a complex expression somewhere. (Lassiter [22] presents a similar view in defense of the brevity-based approach.) This is of course true, but it doesn't explain why a particular context would be such that the more complex expression could be used to begin with. What it shows is that brevity-based approaches must invoke a contextual parameter of "mentionworthiness" that is at least in part independent of considerations of intrinsic brevity or complexity. (Neither author, to my awareness, considers exhaustivity on negative answers like (3).)

Once the need for a contextual "mentionworthiness" parameter is acknowledged, which given example (3) and arguably (5) cannot have anything necessarily to do with intrinsic lexical brevity or complexity, we may as well call the set of propositions that are worth mentioning a "QUD" and get rid of whatever symmetrical notion of relevance was used before (we can always obtain it by closing the notion of mentionworthiness or QUD under negation, should we find a need for it). The resulting picture is essentially what section 2 supports, by explaining why a speaker would choose to address an asymmetrical QUD despite symmetrical interests.

Recall that intrinsic lexical brevity or complexity did not play a role in the explanation I have proposed. Rather, it relied on the obvious brevity benefit of conversationally implicating part of the answer, which obtains in (1) and (3) alike. Another important difference is that, in my explanation, considerations of brevity are not strictly necessary for an audience to be able to identify the exhaustivity implicature: brevity may help explain *why* a speaker chose to address an asymmetrical QUD, but *that* the speaker did so will be evident regardless, from the fact that the utterance would have violated a maxim otherwise (as well as from prosodic focus). In contrast, according to previous brevity-based approaches, the audience would not be able to understand the exhaustivity implicature except through taking brevity into account. That such relatively tiny brevity differences would play such a central role does not seem plausible [5].

4 Conclusion and discussion

This paper argued that even if a speaker's interests are symmetrical – whether in general, which is questionable, or occasionally – it will often be rational for the speaker to organize the propositions of interest into asymmetrical QUDs – and the latter are what matters for exhaustivity. That is, it will be rational to split a symmetrical QUD into two asymmetrical halves because only an asymmetrical QUD permits conveying part of the answer implicitly, i.e., as an exhaustivity implicature, which favors brevity. This is a new explanation for why the alternative sets on which exhaustivity relies tend to be asymmetrical. With previous brevity-based explanations this proposal shares that considerations of brevity have some role to play, though with important differences in the kind of role: the proposed explanation relies only on the general fact that conversational implicature benefits brevity, regardless of the particular lexical entries involved.

I did not discuss the grammatical approach to exhaustivity. Chierchia et al. [7] list the Symmetry Problem as an argument against pragmatic theories of exhaustivity, and in favor of the grammatical approach. This paper shows that this argument falls short: a pragmatic explanation for asymmetrical alternative sets is available. Within the grammatical approach itself, symmetry is sometimes relied upon in order to block certain undesirable exhaustivity inferences. For instance, in order to block the "not all"-inference of the disjunction "some or all", local exhaustification of the first disjunct would render the two disjuncts mutually exclusive, which because of their symmetry would block subsequent global exhaustification (e.g., [6, 20]). Depending on how the sets of alternatives in the grammatical approach relate to

something like relevance or QUDs – and to my awareness there is no concensus in this regard – the arguments in the current paper may have some bearing on the grammatical approach as well. I leave an exploration of this relation to future work.

Zooming out a little, this paper highlighted an important *division of pragmatic labor*, namely between choosing certain QUDs to address and selecting appropriate communicative intents and expressions for doing so. More generally, this is a division between choosing/organizing one's goals and selecting the appropriate means to achieve them. Existing brevity-based approaches to the Symmetry Problem have concentrated on the means, by keeping the symmetrical QUD in place and comparing the brevity benefits only of different ways of addressing that QUD. My proposal, in contrast, considered the brevity benefits of a maneuver at the level of QUDs, i.e., of a certain discourse strategy, and this is what made it both explanatory and more successful at dealing with certain problems. I think it is essential for the field to keep this division of pragmatic labor in mind, and to explicate and motivate assumptions at both levels [32].

5 Acknowledgments

Many thanks to Floris Roelofsen and Jeroen Groenendijk for their comments on many iterations of this work. This project has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 715154). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains.

References

[1] Jay David Atlas and Stephen C Levinson. It-clefts, informativeness and logical form: radical pragmatics (revised standard version). In P. Cole, editor, *Radical pragmatics*, pages 1–62. Academic Press, New York, 1981.

[2] Kent Bach. 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*, pages 21–30. John Benjamins Publishing Company, 2006.

[3] David Beaver and Brady Clark. *Sense and Sensitivity: How Focus Determines Meaning*. Number 12 in Explorations in Semantics. John Wiley & Sons, 2009.

[4] Eliza Block. Is the symmetry problem really a problem? Unpublished manuscript; retrieved from http://web.eecs.umich.edu/~rthomaso/lpw08/block.pdf, 2008.

[5] Robyn Carston. Relevance Theory, Grice and the neo-Griceans: a response to Laurence Horn's 'current issues in neo-Gricean pragmatics'. *Intercultural Pragmatics*, 2:303–319, 2005.

[6] Gennaro Chierchia, Danny Fox, and Benjamin Spector. Hurfords constraint and the theory of scalar implicatures. *Presuppositions and implicatures*, 60:47–62, 2009.

[7] Gennaro Chierchia, Danny Fox, and Benjamin Spector. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. In Claudia Maienborn, Paul Portner, and Klaus von Heusinger, editors, *Semantics: An International Handbook of Natural Language Meaning*, volume 2, pages 2297–2332. Mouton de Gruyter, 2012.

[8] Danny Fox and Roni Katzir. On the characterization of alternatives. *Natural Language Semantics*, 19(1):87–107, 2011.

[9] G. Gazdar. *Pragmatics: Implicature, Presupposition, and Logical Form*. Academic Press, New York, 1979.

[10] Bart Geurts. *Quantity Implicatures*. Cambridge University Press, 2011.

[11] H.P. Grice. *Studies in the Way of Words*. Harvard University Press, 1989.

[12] Jeroen Groenendijk and Martin Stokhof. *Studies on the Semantics of Questions and the Pragmatics of Answers*. PhD thesis, University of Amsterdam, 1984.

[13] Julia Hirschberg. *A Theory of Scalar Implicature*. PhD thesis, University of Pennsylvania, 1985.

[14] Laurence R. Horn. *On the Semantic Properties of Logical Operators in English*. PhD thesis, University of California Los Angeles, 1972.

[15] Laurence R. Horn. Lexical incorporation, implicature, and the least effort hypothesis. In Donka Farkas, Wesley M. Jacobsen, and Karol W. Todrys, editors, *CLS: Papers from the Parasession on the Lexicon*, pages 196–209. Chicago Linguistic Society, 1978.

[16] Laurence R. Horn. Towards a new taxonomy of pragmatic inference: Q-based and R-based implicatures. In D. Schiffrin, editor, *Meaning, Form, and Use in Context*, pages 11–42. Georgetown University Press, 1984.

[17] Laurence R. Horn. *A Natural History of Negation*. David Hume Series on Philosophy and Cognitive Science Reissues. CSLI Publications, Standford, 2001. First published in 1989.

[18] Yan Huang. *Pragmatics*. Oxford Textbooks in Linguistics. Oxford University Press, 2014.

[19] R. Katzir. Structurally-defined alternatives. *Linguistics and Philosophy*, 30(6):669–690, 2007.

[20] Roni Katzir and Raj Singh. Hurford disjunctions: embedded exhaustification and structural economy. In U. Etxeberria, A. Fălăuş, A. Irurtzun, and B. Leferman, editors, *Proceedings of Sinn und Bedeutung 18*. 2013.

[21] Anthony Kroch. Lexical and inferred meanings for some time adverbs. *Quarterly Progress Reports of the Research Laboratory of Electronics*, 104:260–267, 1972.

[22] Daniel Lassiter. Why symmetry is not a problem for a gricean theory of scalar implicature, 2010. Presented at Utterance Interpretation and Cognitive Models 3; retrieved from http://web.stanford.edu/~danlass/Lassiter-SI-UICM.pdf.

[23] Geoffrey Leech. Pragmatics and conversational rhetoric. In H. Parret, M. Sbisa, and J. Verschueren, editors, *Possibilities and limitations of pragmatics*, pages 413–442. John Benjamins, Amsterdam, 1981.

[24] Stephen C Levinson. *Pragmatics*. Cambridge textbooks in linguistics. Cambridge University Press, 1983.

[25] Yo Matsumoto. The conversational conditions on horn scales. *Linguistics and Philosophy*, 18:21–60, 1995.

[26] J.D. McCawley. Conversational implicature and the lexicon. In Peter Cole, editor, *Pragmatics*, number 9 in Syntax and Semantics, pages 245–259. Academic Press, 1978.

[27] Craige Roberts. Information structure in discourse. In J.H. Yoon and A. Kathol, editors, *OSU Working Papers in Linguistics*, volume 49, pages 91–136. Ohio State University, 1996.

[28] Mats Rooth. *Association with Focus*. PhD thesis, University of Massachusetts, Amherst, 1985.

[29] Benjamin Russell. Against grammatical computation of scalar implicatures. *Journal of Semantics*, 23:361–382, 2006.

[30] Westera. An attention-based explanation for some exhaustivity operators. In *Proceedings of Sinn und Bedeutung*. University of Edinburgh, 2017.

[31] Matthijs Westera. 'Attention, I'm violating a maxim!' A unifying account of the final rise. In Raquel Fernández and Amy Isard, editors, *Proceedings of the Seventeenth Workshop on the Semantics and Pragmatics of Dialogue (SemDial 17)*, 2013.

[32] Matthijs Westera. *Exhaustivity and intonation: a unified theory*. PhD thesis, submitted to ILLC, University of Amsterdam, 2017.