Towards a pragmatic explanation for the prevalence of upward-monotonicity in natural language: some results on communicative stability, the strongest answer condition, and exhaustification*

Émile Enguehard
Leibniz-ZAS, Berlin

Abstract

In natural language, logical functions that have the formal property of upward-monotonicity are pervasive. This is most visible in the prevalence of upward-monotonic operators in the logical lexicon. The object of this paper is to point out that this upward-monotonic bias has important consequences for the composition of the alternative sets used in theories of pragmatics, and to argue for the view first proposed by Bar-Lev and Katzir (2022a) that the function of the bias is to guarantee that alternative sets have certain properties. We argue that it is desirable for sets of alternatives to verify the strongest-answer condition, because it is equivalent to the condition that the exhaustification algorithms that have been proposed in the literature to derive scalar implicatures turn the set into a partition of logical space. Then, we show that in certain formal settings, a strongest-answer condition on alternative sets predicts the upward-monotonic bias, assuming those sets are derived as predicted by the structural theory of alternatives. Putting these things together, the bias lets speakers be maximally informative for a given set of alternatives. Because the strongest-answer condition is also equivalent to Bar-Lev and Katzir's stability condition, defined in the context of probabilistic models of pragmatics, our argument can be seen as a generalization of their approach, for which we offer independent motivation.

1 Introduction: monotonicity in natural language

1.1 Outline

The first section of this article reviews how the formal property of (upward-)monotonicity is reflected in natural language, and in particular in the lexicalization patterns of logical operators, which exhibit what we will call an upward-monotonic bias. In the following sections, we will show that the presence of this bias has important consequences for the semantics-pragmatics interface, and, as we will argue, is necessary for certain pragmatics processes,

*I thank Benjamin Spector, Moshe E. Bar-Lev and Roni Katzir for discussion and for encouraging me to write it all, as well as Paula Menéndez-Benito as editor and three anonymous reviewers for *Linguistics and Philosophy* for their very valuable input. All errors are of course mine. This work benefitted from support from the Dutch Research Council (NWO) as part of project 406.18.TW.009 *A Sentence Uttered Makes a World Appear — Natural Language Interpretation as Abductive Model Generation*.

such as exhaustification, to function. In Section 2, we will make a first, informal version of this point: we will see that in fictional versions of English where the upward-monotonic bias does not hold and non-upward-monotonic operators are lexicalized, it could be difficult or impossible to derive scalar implicatures in certain cases, as well as interpret answers to *wh*-questions exhaustively. Section 3 will provide the formal side of the argument: we will define a formal constraint on alternative sets that captures the idea that scalar implicatures should be derived, and we will prove a variety of results to the effect that the upward-monotonic bias is the *only* way to satisfy this constraint. Section 4 concludes the paper by offering a discussion of how the argument improves on previous proposals regarding the upward-monotonic bias, as well as of its potential limitations.

1.2 Basic definitions and examples

Monotonicity Monotonicity is a formal property of Boolean-valued functions defined over ordered sets. A Boolean function can be upward- or downward-monotonic: upward-monotonic functions are such that if any element is mapped to 1, all elements above it are also mapped to 1, while downward-monotonic functions are such that if any element is mapped to 1, all elements below it are also mapped to 1. These properties are rendered formally in (1), for a function named f and an order denoted as $\leq$. Boolean functions can be canonically identified to sets: a set corresponding to an upward-monotonic function is said to be upward-closed, and downward-closed for a downward-monotonic function. Intuitively, one may travel upward in an upward-closed set without leaving the set. For instance, if we have three elements $x_1 < x_2 < x_3$, the sets $\{x_3\}$ or $\{x_2, x_3\}$ are upward-closed, but the set $\{x_2\}$ is not, since going from x_2 to x_3 is an upward move through which we exit the set.[1]

(1) **Upward-monotonicity:** $\forall x, y. [x \leq y \wedge f(x)] \rightarrow f(y)$
Downward-monotonicity: $\forall x, y. [x \leq y \wedge f(y)] \rightarrow f(x)$

The notions of upward- and downward-monotonicity can be applied to the logical words of English and other natural languages, since their denotation is a logical function with arguments of a type that ends in t and is therefore naturally ordered through set inclusion. For instance, the determiner *some* can be analyzed as a function **some** with type $(et)(et)t$ given in (2). The function **some** is upward-monotonic in its second argument (corresponding to the scope of *some*): if **some**$(P)(Q)$ is true, then for any Q' such that $Q \subset Q'$, **some**$(P)(Q')$ is true as well. In this sense, *some* can be seen as an upward-monotonic operator w.r.t. its scope.

(2) **some** $= \lambda P. \lambda Q. \exists x. P(x) \wedge Q(x)$

The analysis of *some* as an upward-monotonic function (w.r.t. its scope) reflects the fact that (3a) entails (3b); replacing the predicate with a superset preserves truth. Similarly, (4b) entails (4a), reflecting the fact that *no* denotes a downward-monotonic function.

[1]The definition offered here is a special case of a more general one: a functions between two ordered sets is upward-monotonic if it preserves the order, and downward-monotonic functions if it reverses it. We get the Boolean case by considering that the set of Booleans is ordered in such a way that $0 \leq 1$.

There are various competing conventions on terminology, depending in part on the domain: instead of upward- and downward-monotonic, one may find *monotone* and *antitone*, *monotonic* and *antitonic*, *order-preserving* and *order-reversing*, or *(monotone) increasing* and *(monotone) decreasing*.

(3) a. Some students speak Polish.
b. Some students speak a Slavic language.

(4) a. No students speak Polish.
b. No students speak a Slavic language.

UE-ness and DE-ness In the semantics literature, the sort of entailment patterns seen in (3) and (4) are often described in slightly different terms: what is diagnosed is that the environment of the object, or of the VP, is *upward-entailing* (UE) and downward-entailing (DE) respectively. UE-ness and DE-ness are directly related to upward- and downward-monotonicity: in particular, argument positions of an unembedded logical operator are upward- or downward-entailing environments if and only if the operator is upward- or downward-monotonic.

The categorization of environments as UE or DE is most famously associated to the distribution of Negative Polarity Items (NPIs): these are elements like English *anybody*, which are illicit in most syntactic environments, but are licensed among other things by negation. Under the Fauconnier-Ladusaw hypothesis (Ladusaw 1980), the distribution of NPIs tracks monotonicity: the environments where they are licensed are exactly the downward-entailing ones.[2]

Convexity While it is not central to the argument, we are going in a few places to use the related notion of convexity. A logical function is convex if its truth set does not have "holes" in it: false inputs that are in-between true inputs with respect to the order. A formal definition is given in (5) for a function f and an order $\leq$.

(5) **Convexity:** $\forall x, y, z. [x \leq y \leq z \wedge f(x) \wedge f(z)] \rightarrow f(y)$

One may easily verify that all monotonic functions are convex. An example of a convex but non-monotonic expression in natural language is *some but not all*, with respect to its scope argument. Since it is non-monotonic, there is no entailment relation between any two of (6a), (6b) and (6c). However, since it is convex, the conjunction of (6a) and (6c) entails (6b). In contrast, *all or none* is non-convex and non-monotonic: (7a) and (7c) can be true while (7b) is false (for instance if the students are half monolingual German speakers and half monolingual Russian speakers).[3]

(6) a. Some but not all students speak Polish.
b. Some but not all students speak a Slavic language.
c. Some but not all students speak a language other than English.

(7) a. All or none of the students speak Polish.
b. All or none of the students speak a Slavic language.
c. All or none of the students speak a language other than English.

[2]Many refinements of this idea have been proposed; we can mention in particular Crnič's (2014) suggestion that the criterion is non-upward-monotonicity rather than downward-monotonicity. The main competing analysis is that of Giannakidou (1998) in terms of veridicality.

[3]This notion is discussed under the name "connectedness" by Chemla et al. (2019) and Enguehard and Chemla (2021). The reason to prefer the name "convexity" is that it is more in line with the established meaning of the words "convex" and "connected" in other domains of mathematics. While there are various ways of filling out the details, a convex set is a set such that everything "in-between" two elements is in the set, while a connected set is a set where there is always a way to go from one element from another.

Some properties of monotonicity and convexity Upward- and downward-monotonicity are both closed under conjunction and disjunction, while negation reverses the direction of monotonicity — here we define conjunction, disjunction and negation over all types that end in t in the natural way. This is a specific instance of the fact that upward- and downward-monotonicity are closed under composition with upward-monotonic functions, while composition with downward-monotonic functions reverses the direction, combined with the fact that conjunction and disjunction are themselves upward-monotonic (with respect to both arguments), while negation is downward-monotonic. In particular, going back to natural language, any expression of the form "not OP", where "OP" is upward-monotonic, is downward-monotonic, and as we will see, many downward-monotonic expressions have this form (e.g. *not all*).

As already mentioned, all monotonic functions are convex. Convexity is not closed under negation in general, nor under disjunction, but it is closed under conjunction. The conjunction of an upward-monotonic and a downward-monotonic function is convex, and in natural languages many non-monotonic expressions have this form (e.g. *some but not all*).

1.3 Monotonicity in the lexicon

A remarkable property of natural language is that lexicalized logical operators are predominantly upward-monotonic, a fact we will call the upward-monotonic bias of natural language. Table 1 lists the monotonicity status of a variety of English logical words: we can see that all listed single-word operators are monotonic, and that most of them are upward-monotonic. Furthermore, a majority of downward-monotonic operators are transparently derived from their upward-monotonic negation through combination with a negative morpheme: *never* from *ever*, *either... or* from *neither... nor*, *impossible* from *possible*, etc. We can also note that there are no downward-monotonic operators the negation of which is not lexicalized as well, while some upward-monotonic operators do not have lexicalized negations.

These observations are well-established in the literature. An early generalization in terms of monotonicity is found in Barwise and Cooper's (1981) universal U5, which states that if a downward-monotonic quantifier is expressible, so is its upward-monotonic negation, as we have just noted.[5] Horn (1989, see especially chap. 3 and 4) provides an extensive survey of the arguments in favour of the idea that negativity is marked: on top of the morphological evidence, which extends widely across languages, it is noted in particular that negative operators trigger more specific pragmatic inferences compared to their positive cousins, and sentences containing them are more difficult to process. While the notion of negativity discussed by Horn extends beyond purely logical words, and is defined through linguistic tests and intuitions rather than semantically, it coincides exactly with downward-

[4]It is sometimes said that English *nor* on its own expresses negated disjunction (logical NOR), which would warrant its inclusion in the table as a downward-monotonic connective. However, attested examples where *nor* is used without *neither* and without a preceding negation are exceptional (see Horn 2012 for some of them). *nor* and its counterparts in other languages are more readily analyzed either as an asymmetric presuppositional conjunction (roughly *and not either*), as defended by Wurmbrand (2008) for instance, or as a disjunction which is also some kind of strong NPI or N-word (see for instance Gajić 2019).

[5]Barwise and Cooper's U5 bears on "simple" DPs, which they define to include one-word quantifiers, DPs formed from quantifying determiners, and modified numerals, but not DPs containing overt negation and connectives. In this respect it differs from our generalization which bears on single-word operators of any category. Their universal U6 is also interesting to us: it entails that all simple expressions have convex meanings. See also Keenan and Westerståhl (1997) for similar claims.

Class	**Upward-M**	**Downward-M**	**Non-M**
Quantifiers	*somebody*	*nobody*	
	everybody	(*not everybody*)	
Determiners	*some*	*no*	(*some but not all*)
	all (scope)	(*not all*)	
	ten	(*fewer than ten*)	(*exactly ten*)
	many	*few*	
	most (scope)		
Connectives	*or*	(*nor*[4])	(**xor*)
	and	(**nand*)	
Complex connectives	*either... or*	*neither... nor*	
	both... and	*(not both... and)*	
Temporal adverbs	*sometimes/ever*	*never*	
	often	*rarely*	
	always	(*not always*)	
Modals	*may/can*	(*cannot*)	(*may or may not*)
	must	(*needn't*)	
Modal adjectives	*possible*	*impossible*	
	necessary	*unnecessary*	

Table 1: The monotonicity properties of various English logical words. Some multi-word expressions are included to give an idea of what meanings are "missing". When several expressions are on the same line, the second column is equivalent to the negation of the first, and the third column is equivalent to the conjunction of the expression in the first column with the negation of a stronger scalemate. Our classification assumes that the "at least" reading of numerals is basic (cf. discussion in Spector 2013) and does not take into account the restrictor argument of determiners — cf. Section 4.3 for discussion.

monotonicity as far as the logical lexicon is concerned.

One particular framing of the upward-monotonic bias that has received a lot of attention in the literature is the "missing O-corner" problem identified by Horn (1973). As long as they are monotonic and "at the end of the scale", that is to say, either strongest or weakest among their alternatives, logical operators can be classified into four categories, known as the corners of the Square of Opposition. For instance, in the case of partitive determiners, the four corners are *all* (strong upward-monotonic), *none* (strong downward-monotonic), *some* (weak upward-monotonic) and *not all* (weak downward-monotonic). Using these four items as representatives of their respective corner, Horn (1973) observes the following lexicalization hierarchy: *all*/*some* > *none* > *not all*. Across languages and categories, elements lower in the hierarchy are less likely to lexicalize and are almost always more complex morphologically than elements further up. In particular, the weak downward-monotonic category, also known as the O-corner, almost never lexicalizes — a common way of summarizing this observation is to note that English lacks **nall* and **nand*. Horn (1973; 1989) offers a pragmatic explanation for this latter fact, the second inequality in the hierarchy: once the other three corners are lexicalized, the O-corner is dispensible in a certain sense. However, this explanation is compatible with unattested lexicalization patterns; to rule those out, Horn appeals to the inherent markedness of negation (or downward-monotonicity). Thus the first inequality in the hierarchy, which is essentially the upward-monotonic bias, is assumed to have deeper roots. The argument is made more precise by Katzir and Singh (2013), who propose that logical words are represented at some level in a certain logical language with upward-monotonic primitives, so that downward-monotonic or non-monotonic operators are necessarily more complex.

Katzir and Singh's account of the O-corner problem predicts the upward-monotonic bias; in fact, it makes the stronger prediction that all simplex operators are upward-monotonic, and that all downward-monotonic operators are derived from upward-monotonic ones.[6] We already noted that there are only a few downward-monotonic operators that are not transparently derived from an upward-monotonic operator with an added negation; the stronger characterization of the bias extends this decomposition to those few exceptions, essentially arguing that *few* is underlyingly *not many*, *nobody* is *not anybody*, etc. In addition to the O-corner problem, the strong version of the bias helps explain the general pattern of negative markedness as well as the fact, noted above, that lexical downward-monotonic operators never come without their upward-monotonic negation.[7] As Katzir and Singh (2013) discuss, there is also direct evidence for the separation of negation from various negative operators: it has been used to account for split-scope readings and for negative concord in particular.[8] Finally, as we are going to review in detail in the next section, this assumption is necessary to explain certain patterns of implicature derivation under a structural theory of alternatives. For all these reasons, we will adopt here the strong characterization of the upward-monotonic bias, and this is what we will be trying to explain.[9] We can remark that our characterization of the bias also applies to negation itself: the upward-monotonic "operator" corresponding to the negation of negation is the absence of any operator, which is less complex than *not* in a fairly indisputable way.

The fact that downward-monotonic operators are simpler than non-monotonic ones can be seen as a direct consequence of the upward-monotonic bias: if we have a lexicon of basic upward-monotonic operators, we can construct downward-monotonic ones by composing the upward operators with negation. No other class of operators can be obtained in such a simple way, since negation is the only non-trivial type $\langle t, t\rangle$ function. For this reason, there is no need to explain the prevalence of downward-monotonic operators in a separate way.[10]

[6]The way we formulated the claim entails that all monomorphemic operators are upward-monotonic, but it has the advantage of also saying something about complex upward-monotonic operators: we capture the fact that (on the evidence of split-scope readings, as seen below) *few* is more complex than *many*, independently of whether *many* is analyzed as simplex or not.

[7]In Section 4.2, we will discuss in which sense Katzir and Singh's (2013) proposal and other recent approaches to the O-corner problem explain the upward-monotonic bias, and in which sense they fail to do so.

[8]See for instance Zeijlstra 2004 on negative concord, Jacobs 1980 on split-scope and Penka 2010 on both. See also Abels and Martí 2010 for a non-decompositional account of split-scope. Some of these authors (Penka (2010) in particular) in fact defend that at least some negative operators are upward-monotonic operators that co-occur systematically with sentential negation.

[9]We will assume throughout the discussion that non-upward-monotonic operators are more complex at the level of morphosyntax; this is for the sake of concreteness and our argument could also go through if we locate the greater complexity at the level of cognitive representations. See also footnote 38.

[10]The case of convexity is not as clear. From a basic lexicon of upward-monotonic functions, including conjunction and disjunction, plus negation, many convex but not monotonic functions can be derived as $F \wedge \neg G$; for instance *exactly one* is equivalent to *one and not two*. Many non-convex functions can be derived in an equally simple way, as $F \vee \neg G$ — for instance, *many or not any*. In this sense, the preference for convex functions does not follow from the upward-monotonic bias. It would possibly follow if we added the claim that conjunction is more basic than disjunction, or that exhaustification (which generates propositions of the form $F \wedge \neg G$) is a primitive.

1.4 Towards an explanation of the upward-monotonic bias

Before we move on to the main argument of the paper, let us briefly touch on why it is that the upward-monotonic bias is in need of an explanation. It is easy to imagine a language that is just like English, but lexicalizes **nall*, **nand* and other O-corner operators in addition to or instead of English's actual operators. Speakers do produce sentences that involve O-corner operations and would be simpler if O-corner operators were lexicalized.[11] Thus O-corner operators would make communication more efficient in some sense. We can also imagine that non-monotonic operators like **xor* or **justone* could be lexicalized — these are useful enough that they are often defined in formal languages.

One may appeal to some kind of pressure towards economy to explain that not all possible operators are lexicalized: since we have negation and binary connectives, arbitrary logical functions can be expressed in combinatorial ways, and are not needed as lexical items. However, this would not explain why it is specifically upward-monotonic operators that are lexicalized, as other choices could let us be as expressive. Furthermore, in other areas of the lexicon, there does not seem to be a general principle that what can already be expressed in combinatorial ways should not be lexicalized: in English, lexical items like *twelve* (**twenteen* or simply **ten-two*) or *son* (*male child*) clearly run afoul of such a principle.

For these reasons, the upward-monotonic bias in the lexicon is not a necessary feature of natural language: we need to explain why non-upward-monotonic operators are always derived, and that means both why we do not see basic downward-monotonic operators and why we do not see basic non-monotonic operators — as we will see, some earlier explanations fail to address both. In the rest of this paper, we will offer an explanation that is rooted in implicature derivation patterns: the claim will essentially be that the upward-monotonic bias is the only way to make implicature derivation work in desirable ways. This is because certain logical relations need to hold between the alternatives involved in implicature derivation, and the upward-monotonicity of basic sentences necessitates the upward-monotonicity of more complex ones and therefore of the lexicon.

The approach of explaining the upward-monotonic bias through constraints on the structure of alternative sets has recently been proposed by Bar-Lev and Katzir (2022a). They show that, in the case of binary connectives, the prevalence of upward-monotonic operators (conjunction and disjunction) can be explained if alternative sets have to have a certain property called communicative stability, defined in the context of probabilistic models of pragmatics. Our proposal is the same at the formal level, in the sense that the property we are going to propose to enforce on alternative sets is equivalent to Bar-Lev and Katzir's communicative stability, and we will call it stability as well for that reason. However, we are going to prove the applicability of the approach to more cases, and to offer a novel motivation for stability, without reference to probabilistic models. Other approaches to explaining the upward-monotonic bias will be discussed in Section 4.2; we will argue that they fail at really explaining as adequately as the stability approach does.

[11] See Section 4.3 for a discussion of the use conditions of O-corner statements.

2 The upward-monotonic bias and the semantics/pragmatics interface

We have seen that upward-monotonic logical operators are prevalent in natural language. In this section, we will argue that various pragmatic processes depend on this property. The general idea is that the processes we are going to discuss can be described with reference to certain alternative sets consistent with the structural theory of implicatures (Katzir 2007; Fox and Katzir 2011), but without the upward-monotonic bias, there would be no way of deriving the appropriate sets of alternatives.

2.1 The symmetry problem in the derivation of implicatures: the direct case

The utterance of (8a) is generally interpreted as implying that (8b) is false. This has been categorized as a case of a scalar implicature, caused by the use of a scalar element, *some*, that has a salient alternative *all*. The established explanation goes like this: if the speaker believed (8b) to be true, saying (8b) would have been more informative; the speaker did not say (8b) and therefore does not believe (8b) to be true. One may either view this inference as proceeding from general-purpose reasoning or as a grammaticalized process (Sauerland 2012).

(8) a. Mary ate some of the cookies.
    b. Mary ate all of the cookies.

One problem with the above theory is that if we apply the same thinking to (8c), which is also more informative than (8a), we predict that (8a) should lead to the inference that (9a) is false, and therefore, that (8b) is true; this is never observed. Hence, while (8b) can be an alternative of (8a), (9a) cannot. On top of that, there are independent reasons to think that sentences which are logically independent from the utterance can lead to implicatures (see for instance Fox 2007) — throughout this discussion, we will assume that any non-weaker alternative can potentially be excluded to lead to an implicature. Taking this into account, we might expect to see the inference that (9b) is false from the utterance of (8a). This would again lead to the conclusion that (8b) is true; since that is never observed, (9b) also cannot be an alternative. The challenge of explaining this discrepancy between (8b) on the one hand and (9a) and (9b) on the other hand is known as the symmetry problem (Breheny et al. 2018). The name comes from the fact that (8b) and (9a) are logically symmetric with respect to (8a): both (8b) and (9a) asymmetrically entail (8a), and there is no purely logical way of distinguishing them in relation to (8a). Theories of implicature derivation (such as Sauerland 2004) predict that when symmetric alternatives are present, we should observe ignorance inferences rather than scalar implicatures.

(9) a. Mary ate some but not all of the cookies.
    b. Mary did not eat all of the cookies.

Fox and Katzir (2011), building up on Katzir (2007), propose that alternatives are obtained by replacing elements of the utterance by alternative elements of equal or lower syntactic complexity; this proposal is known as the structural theory of alternatives. Since (9a) involves a coordinated structure, and both (9a) and (9b) involve negation, they are more complex than (8a) and (8b) with their one-word determiner in the object position.

This explains the absence of (9a) and (9b) among possible alternatives. Conceptually, one way of justifying this line of explanation is to relate it to Grice's (1975) Maxim of Manner, *Be Brief*: the reason the speaker who uttered (8a) did not use (9a) or (9b) might simply be that they were too long, so we should not conclude anything about their communicative intentions from this choice.[12]

What is of interest to us here is that this explanation is dependent on the fact that no simple way to express (9c) and (9d) exists. Thus it is the upward-monotonic bias which breaks the symmetry: because (9c) and (9d) involve non-upward-monotonic quantifiers, and those are dispreferred in the lexicon, they are more complex to express and do not interfere with pragmatic reasoning. In a fictional version of English where non-upward-monotonic operators are lexicalized and as simple as upward-monotonic ones, (10a) and (10b) could be alternatives to (9a) — here and below ★ indicates sentences in this fictional English with no lexical restriction. Then, depending on one's exact theory of implicature derivation, (9a) would trigger the obligatory inference that the speaker is ignorant, or it could not be used without contradiction, or it would lead to no implicature at all, or it could have contradictory implicatures depending on the context. Of course, a speaker wishing to express the enriched meaning of (8a) in this fictional English could simply use (10a), which encodes it lexically. As we are going to see however, the interaction between fictional non-monotonic operators and negation causes problems of its own.[13]

(10) a. ★ Mary ate some-but-not-all of the cookies.
b. ★ Mary ate not-all of the cookies.

Note that while we have been using partitive determiners as an example here, the point can be made with the other categories of logical words mentioned in the table as well. Considering for instance binary connectives, in (11), the (a) sentence implicates the falsity of the (b) sentence, which we are able to explain through competition and the structural theory because the (c) and (d) sentences are not possible in English.

(11) a. Mary talked to John or Peter.
b. Mary talked to John and Peter.
c. ★ Mary talked to John xor Peter.
d. ★ Mary talked to John nand Peter.

[12] This Gricean motivation for the structural theory is found in Katzir 2007, but not in Fox and Katzir 2011.

[13] Another problem with this sort of non-monotonic operators is that they compose poorly. The combination of *some-but-not-all* with itself or with *some*, as in (ia), (ib) and (ic), expresses a non-convex function of the predicates denoted by the content words in the sentence. Such meanings have been argued to cause cognitive difficulties (Chemla et al. 2019; Enguehard and Chemla 2021). Meanwhile, Bassi et al. (2021) argue in favour of two possible enriched meanings for (id), both of which are convex and cannot be expressed easily using *some-but-not-all*. Thus *some-but-not-all* is not that useful for expressing what natural language sentences actually convey beyond the simplest cases. This point connects to the formal discussion in Section 3.2.6.

(i) a. ★ Some-but-not-all of the students did some-but-not-all of the exercises.
b. ★ Some of the students did some-but-not-all of the exercises.
c. ★ Some-but-not-all of the students did some of the exercises.
d. Some of the students did some of the exercises.

2.2 The case of indirect implicatures

Implicatures can also arise when using scalar items with weaker alternatives in downward-entailing contexts. These cases are have been called indirect implicatures by Chierchia (2004). One typical case is that of (12a): with the appropriate prosody, as suggested by capitalization, it leads to the inference that (12b) is false, suggesting that (12b) is among its formal alternatives.[14]

(12) a. Mary didn't eat ALL of the cookies.
b. Mary didn't eat many of the cookies.

As Breheny et al. (2018) discuss, the case of indirect implicatures is potentially problematic for Fox and Katzir's theory of alternatives. Once again, there is a fictional *many-but-not-all* alternative (13a) that could lead to an instance of the symmetry problem, but it does not exist due to the upward-monotonic bias. The problem lies with the *many* alternative (13b), which is simpler than the postulated alternative (12b) — it can be obtained from it through deletion of *not* — and is not entailed by the utterance. Negating this alternative should lead to the inference that (12b) is true, contrary to fact. Thus in this case, the alternative set appears to be subject to more constraints than what the structural theory predicts.

(13) a. ★ Mary ate many-but-not-all of the cookies.
b. Mary ate many of the cookies.

As a simple descriptive solution to this problem, we will adopt here the assumption that negation may not be deleted to form alternatives. Under this assumption, the competitors to (12a) cannot include (8a) or (13b). We could however still run into the symmetry problem if downward-monotonic simplex alternatives were available. In particular, (14a) would be symmetric with (12b). The upward-monotonic bias avoids this situation by ensuring that *few* is more complex than *many*, so that (14a) needs not be in the alternative set when (12b) is — here we see how, once we accept the structural theory, the absence of symmetry in indirect implicatures is evidence for the claim that *few* or *none* are complex. Note that in this configuration, the sort of non-monotonic operators that we considered in the previous section would also be problematic: for instance, (14b) would also be symmetric with (12b).

(14) a. Mary didn't eat few of the cookies.
b. ★ Mary didn't eat some-but-not-many of the cookies.

Once again, a similar point can be made when considering binary connectives. The sentence in (16a) lead to the inference that (16b) is false;[15] we can explain this in our amended

[14] The indicated prosody is necessary for this inference to obtain; with a more neutral delivery, the inference we observe is only that Mary ate some of the cookies. The weaker inference can also be analyzed as an indirect implicature due to the *not some/any* alternative, but we focus on the strong inference here as it creates a configuration where the presence of non-monotonic operators in the lexicon will turn out to be problematic, and this provides an illustration of our later formal developments.

Another complication that we set aside here is that the exact inference that (12a) triggers is vague and might well be different from the negation of (12b); this is unclear as (12b) is itself vague. Again, here our purpose is to illustrate a certain formal configuration; the exact identity of the quantifier from which the inference is derived (*many*, *most*, *nearly all*...) does not affect the argument. The important thing is that the quantifier *many* stands for can compete with *all* in some cases.

[15] Emphasis on *and* is required to bring out this reading rather than a homogeneous reading where Mary didn't talk to either person.

structural theory if (16b) is an alternative to (16a), but our explanation depends of the non-existence of (16c) in order to avoid symmetry.[16]

(15) a. Mary won't talk to John AND Peter.
b. Mary won't talk to John or Peter.
c. ★ Mary won't talk to John nor Peter.

Earlier, we noted that speakers of fictional English might not find it a problem that implicature derivation in direct cases does not obtain, given that they can express the same meanings lexically. In principle, there could also be lexicalized operators that express the desired meanings under negation. To express the enriched meaning of (12a) using a negation, one would need a lexicalized version of *either few or all*. This expression is not just non-monotonic, but also non-convex, and there is some evidence that such logical functions are cognitively hard to process (Chemla et al. 2019). Additionally, such functions do not correspond to any obvious natural mathematical class, so that if they are lexicalized, there is no limit on what can be lexicalized at all, which is arguably unrealistic.[17]

2.3 Partial excursum on the semantics of questions

The final item in this review of the interaction between the upward-monotonic bias and pragmatics will look at how the alternatives forming question meanings also appear to be influenced by the upward-monotonic bias.

To begin with, in a question-answer pair like (16), an inference that the given positive answer is an exhaustive description of the witnesses to the *wh*-word is generally observed. This inference can straightforwardly be analyzed as a scalar implicature due to either (17a) or (17b) being a competitor. The alternative set from which that implicature is derived cannot contain negative propositions like (17c), as otherwise there would be a situation of symmetry. The structural theory explains this straightforwardly, since (17c) is more complex than the utterance, which reflects the fact that negation is marked.

(16) Q: Who was there?
A: Mary was.
⇝ John was not there.

(17) a. John was there.
b. John and Mary were there.
c. John wasn't there.

Because of this exhaustivity implicature, in a situation where there are several witnesses (several people were there), one cannot use a simple positive sentence like (17a). Instead, it is necessary to use a conjunctive sentence like (17b). This is another case where the attested lexicon of logical operators meshes well with pragmatic processes: here, the absence of conjunction in language would be problematic in the presence of implicatures, in contrast with the earlier cases we discussed, where the presence of various operators would be

[16]Of course (15c) does exist in English, but it does not mean what it should mean if *nor* was the lexicalization of logical NOR; instead it is synonymous with (15b). In our fictional English, (15c) is synonymous with (11a).

[17]These points are taken up again in a more formal way in Sections 3.2.5 and 3.2.6, where we will see that a restriction to upward-monotonic operators is the only way to derive strong implicatures without non-convex propositions or while keeping to natural mathematical classes of functions respectively.

problematic.

In the answer set theory of questions, questions denotes sets of propositions, interpreted as potential answers. The answers included in the answer set are exactly those that can be interpreted as complete answers. Thus, propositions equivalent to conjunctive answers like (17b) are part of the answer set.[18] Meanwhile, negative propositions like (17c), which are not generally judged to be complete answers, are not in the set. The resulting sets verify a structural condition known as the strongest true answer condition: in every possible situation, one of the propositions in the answer set is the single strongest true member of the set. It has been claimed that questions have to obey this condition to be well-formed, for reasons that are beyond the scope of this paper (Dayal 1996).

Note that we can explain the ignorance inferences triggered by negative answers within our theories of implicature derivation by assuming that the set of alternatives involved in the interpretation of negative answers contains both positive and negative answers. This set of alternatives is exactly what the structural theory predicts, as long as deleting negation is allowed.[19] In such a set, every alternative has a symmetric partner, namely its negation: excluding ψ makes it impossible to exclude $\neg\psi$, and vice-versa. According to theories of implicature derivation, this situation prevents strengthening, and leads to ignorance inferences. Regardless of what happens in the case of negative answers, the set of alternatives used in the interpretation of positive answers, including conjunctive ones, is correctly predicted by the structural theory not to include negative answers, which lets us avoid symmetry and interpret them as complete answers. As we have seen in the previous section, this prediction is dependent on the absence of downward-monotonic connectives.

Let us recapitulate: the answer sets of questions have to include conjunctions but not negations to verify the strongest true answer condition. The alternative sets used in the interpretation of answers also have to exclude negations to avoid symmetry (whether they contain conjunctions depends on the specific utterance). In the next section of this paper, we are going to see that this is not a coincidence, as the strongest true answer condition and the condition that strong implicatures can be derived are essentially equivalent. Thus, the argument will go, the upward-monotonic bias in the lexicon comes from the need to have alternative sets that verify the strongest true answer condition, which in turns allows for strong implicatures.

While this assumption is not essential to our argument, the similarity between answer sets and sets of alternatives makes it tempting to conclude that they are derived in similar way, that is, that the derivation of answer sets is constrained by structural complexity. If that is so, then the make-up of answer sets is also dependent on the upward-monotonic bias: it is because conjunction is lexicalized and downward-monotonic binary connectives are not that answer sets have the shape that they have, and that they can verify the strongest true answer condition. The general argument of the paper would then apply to both alternative sets and answer sets. Evidence for this assumption can be found in the cases discussed by Spector (2008) of questions where the *wh*-word ranges over higher-type elements, but

[18] Note that uses of *and* between type *e* expressions is often assumed to correspond to sum/set formation and not Boolean conjunction, which does not affect the truth-conditional effect in simple cases.

[19] Thus, it could be that the ban on deleting negation observed in the case of indirect implicatures follows from an alignment of factors that fails to obtain in the case of question-answer pairs. It could also be that negation can never be deleted, but that in question-answer pairs, the positive answers, because they form the answer set, are always included in the set of alternatives along with any structurally-derived alternatives. See also Spector 2007 for a discussion of non-positive answers to questions and of their implicatures; Spector's account is not based on a structural view of alternatives and instead assumes certain semantic restrictions on the alternatives of non-positive answers.

with a restriction to upward-monotonic ones. The essential data point is that in a context where we have established positive and negative quantified obligations, ascribing knowledge about these obligations using a simple *wh*-phrase seems to only entail that the attitude holder knows about the positive obligations, and not that they know about the negative ones. Concretely, in the specified context in (18), (18a) appears to entail (18b) but to be compatible with the falsity of (18c). This is readily explained if the *wh*-phrase in (18) ranges over upward-monotonic quantifiers only.[20] This constraint, in turn, can be explained in the spirit of the structural theory by virtue of the upward-monotonic bias: the *wh*-word ranges over quantifiers up to a certain level of complexity, which are in fact the upward-monotonic ones.

(18) Context: *Jack must read* The Idiot *or* Crime and Punishment*, whichever he prefers, and is not allowed to read* The Brothers Karamazov.

a. Sue knows which novels Jack must read.
b. Sue knows that Jack must read *The Idiot* or *Crime Punishment*.
c. Sue knows that Jack must not read *The Brothers Karamazov*.

3 Formal results: linking exhaustification, the strongest true answer condition and upward-monotonicity

In the previous section, we have shown than in a variety of cases, deviating from the actual state of the lexicon, where the simplest operators are the upward-monotonic ones, would interfere with the derivation of scalar implicatures. This paves the way for a pragmatic functionalist explanation of the upward-monotonic bias: it evolved in language to let speakers make the most of pragmatic processes, so that they could be maximally informative through implicatures all while using a restricted lexicon and simple utterances. In this section, we will establish further support for this argument by going through a series of formal results to the effect that the lexicalizing upward-monotonic operators is in fact the *only* way, assuming various plausible additional constraints, to derive sets of alternatives with the appropriate structure.

We will begin by showing that the strongest true answer condition, and the condition that exhaustification of the set of alternatives should result in a partition of logical space — which is how we can formalize the desideratum of deriving implicatures — are equivalent. We will call alternative sets that verify these equivalent conditions *stable*, because the conditions in question are also equivalent to the property of communicative stability defined by Bar-Lev and Katzir (2022a), as detailed in Section 3.3. Then, we will show that in various formal settings, a lexicon of upward-monotonic operators is the only way to make it so that the sets of alternatives derived from the structural theory are stable, beyond the specific case of symmetric binary connectives discussed by Bar-Lev and Katzir.

We are going to focus on an idealized situation where a certain set of alternatives A is clearly identified as being the set of competitors. The idea is that A contains the denotations of an alternative set obtained through structural derivations in the spirit of Fox and Katzir (2011): it includes all propositions expressible using a certain structure or simplifications of it. Because the derivation of A involves considerations of complexity, its contents will be influenced by lexicalization facts.

[20] See also Fox 2020.

3.1 Formalizing stability: the strongest true answer condition and exhaustification

The derivation of implicatures can be formalized with the use of exhaustification operators. An exhaustification operator is a function of two arguments: if φ is a proposition in our set A, $\textsc{exh}_A(\varphi)$ is the enriched meaning of an utterance whose literal meaning is φ and whose competitors' literal meanings are gathered in A. The operation represented by EXH is thought to be roughly similar to the meaning of *only*; a simple definition for it is that offered by Krifka (1993) for *only*, given here in (19). Under that definition, EXH conjoins the utterance with the negation of all non-weaker alternatives.

(19) $\textsc{exh}_A(\varphi) := \varphi \wedge \left(\bigwedge_{\psi \in \mathrm{NW}_A(\varphi)} \neg\psi\right)$
where: $\mathrm{NW}_A(\varphi) := \{\psi \in A : \varphi \not\models \psi\}$

A specificity of this definition is that in the presence of symmetric alternatives, the exhaustified proposition is a contradiction.[21] Other, more sophisticated definitions have been proposed to avoid this property; for instance, the "innocent exclusion" operator (Fox 2007) does not derive implicatures from symmetric alternatives. The facts presented in this section apply to the basic definition as well as the innocent exclusion operator and the "minimal world" operator (Schulz and van Rooij 2006; Spector 2006), but not to the "innocent exclusion and inclusion" operator (Bar-Lev and Fox 2020); those other operators are defined and discussed in more detail in Appendix A.1.

From the set of alternatives A, we can define the set of exhaustified meanings $\textsc{exh}(A)$, defined in (20): $\textsc{exh}(A)$ contains the exhaustification against A of every proposition in A. The propositions in $\textsc{exh}(A)$ correspond to what speakers can communicate, taking implicatures into account, in an idealized situation where the set of competitors is clearly identified as being A. In principle, there is a risk of not gaining or even losing expressivity when going from A to $\textsc{exh}(A)$: for instance, it might be the case that $\textsc{exh}(A)$ does not contain any propositions that are true in certain situations even though A does. Here we are going to assume that the optimal situation is one where $\textsc{exh}(A)$ contains a subset of propositions that cover A — there is always something true — and are as informative as something derived from the propositions in A can be.

(20) **Exhaustified set:** $\textsc{exh}(A) := \{\textsc{exh}_A(\varphi) : \varphi \in A\}$

We formalize this latter condition of maximal informativity through the notion of the partition induced by A. For any possible world w, we can define the true set $\mathbb{T}_A(w)$ of propositions true at w, and the false set $\mathbb{F}_A(w)$ of propositions false at w, as per (21). The propositions obtained by conjoining all propositions in $\mathbb{T}_A(w)$ and the negations of all propositions in $\mathbb{F}_A(w)$, for all w, form a partition of the set of possible worlds which we call the partition induced by A, $\mathcal{P}(A)$, defined in (22).[22] If two worlds are in the same cell of the induced partition, all propositions in A have the same truth value in them, so that it is not possi-

[21] Here we define a set of alternatives as symmetric when the disjunction of all elements of the set is entailed by the utterance, and the set is minimal for this property.

[22] Proof: at any world w, the proposition of $\mathcal{P}(A)$ formed from the true and false sets of w is clearly true, so $\mathcal{P}(A)$ covers the space of posible worlds. If two worlds have the same true and false sets, they contribute the same proposition to $\mathcal{P}(A)$; if they have different true and false sets, there is φ that is true at one and false at the other, and the propositions contributed by the worlds respectively entail φ and $\neg\varphi$, which means they are disjoint. Equivalently, $\mathcal{P}(A)$ can be seen as the partition associated to the equivalence relation over worlds $\sim_A$ defined by: $w \sim_A w'$ if and only if $\mathbb{T}_A(w) = \mathbb{T}_A(w')$.

ble to distinguish them using logical combinations of elements of A. Thus, the cells of the induced partition correspond to the maximal degree of informativity achievable using A, assuming arbitrary logical combinations are allowed.

(21) **True set:** $\mathbb{T}_A(w) := \{\varphi \in A : \varphi(w)\}$
False set: $\mathbb{F}_A(w) := \{\varphi \in A : \neg\varphi(w)\}$

(22) **Induced partition:** $\mathcal{P}(A) := \left\{\left(\bigwedge_{\varphi\in\mathbb{T}_A(w)} \varphi\right) \wedge \left(\bigwedge_{\varphi\in\mathbb{F}_A(w)} \neg\varphi\right) : w\right\}$

As a simple example, consider the set of propositions A_0 in (23a). The induced partition of A_0 has four cells, corresponding to the propositions in (24). In general, the partition induced by n propositions has at most 2^n cells. If the propositions are logically related, the induced partition can be smaller: for instance, the induced partition of A_1 in (23b) is the same as that of A_0.

(23) a. $A_0 := \{\text{John is here } (p), \text{Mary is here } (q)\}$
b. $A_1 := \{p, q, p \vee q\} = \{\text{John is here, Mary is here, John or Mary is here}\}$

(24) $\mathcal{P}(A_0) = \mathcal{P}(A_1) = \{\neg p \wedge \neg q$: Neither John nor Mary is here
$p \wedge \neg q$: John is here but Mary is not,
$\neg p \wedge q$: John is not here but Mary is,
$p \wedge q$: John and Mary are both here$\}$

In principle, as the example in (24) illustrates, speakers can express cells of the induced partition using logical combinations of propositions in A with logical connectives like conjunction, disjunction and negation. However, this presumably involves using more complex sentences than the ones that express meanings in A. Ideally, exhaustification lets them express the partition cells using the simpler sentences whose meanings are in A: in this way, they are both maximally informative and concise. For this reason, we will consider that it is desirable that EXH(A) should contain the induced partition.[23]

The theorem this section is leading to is that the condition we just stated, that EXH(A) should contain the induced partition is equivalent to the strongest true answer condition

[23] If it is possible for all propositions in A to be false at the same time, there will be a corresponding cell in the induced partition where no proposition in either A or EXH(A) is true. Our condition therefore requires A to cover logical space: there should be true propositions at every world. As a reviewer points out, this is often not true of alternative sets found in the literature, like $\{p, q, p \vee q, p \wedge q\}$ (as typically assumed, the alternatives of a declarative disjunction) or $\{p, q, p \wedge q\}$ (as typically assumed, the denotation of a constituent question over a two-element domain). Compared to such sets, the sets we are going to consider in this paper have extra tautologies or negative propositions (e.g. $\neg p \wedge \neg q$) that ensure they cover logical space. We could avoid this discrepancy by relativizing the condition to the set of worlds covered by A; all the results would be preserved, and would apply to "naturalistic" alternative sets. The reason we do not do so is to simplify the formal statements.

As the reviewer also points out, in principle, it would be more principled for the condition to be that EXH(A) should partition out the context set in a given conversational situation, and not logical space as a whole. This condition is, in principle, weaker, but we can conjecture that the only way to guarantee that it is met in every possible situation is for the condition over logical space to hold.

Finally, note that if we started from this idea that EXH(A) should partition out the context set, and also assumed that A always covers the context set in a given conversational situation, the natural sufficient condition we would arrive at is the condition we mooted above, that is, the requirement that EXH(A) should partition out the set covered by A. The assumption that A covers the context set is often made in the context of question semantics — it predicts that constituent questions presuppose the existence of witnesses, cf. Dayal (1996) — but is less systematic in the literature on implicature derivation and focus semantics.

— which will be called STA from now on — applied to A. Furthermore, we can identify the subset of A from which the induced partition is derived through exhaustification: it is specifically the propositions that do not have a stronger alternative everywhere where they are true, which we are going to call the non-dominated propositions. For instance, in A_1 as defined above, $p \vee q$ is dominated since whenever it is true, at least one of p or q is also true. Note that a well-informed speaker who obeys the Maxim of Quantity should never used dominated propositions, since there is always a better alternative. The formal definition of domination is given in (25), allowing us to then state our result in (26). When an alternative set verifies the three equivalent conditions in (26), we will call it *strongly stable*.

(25) **Domination:** for a set of propositions A and an element of this set φ, φ is dominated in A if there is a subset S of A such that:

a. all elements of S asymmetrically entail φ;
b. φ entails the disjunction of all elements of S.

(Note that a contradiction is always dominated, taking S to be the empty set and relying on the usual convention that an empty disjunction is a contradiction.)

(26) **Theorem:** for a set of propositions A, let A^* be the set of non-dominated alternatives in A. The following conditions are equivalent:

a. For every w, $\mathbb{T}_A(w)$ has a minimal element under entailment (STA).
b. $\text{EXH}(A)$ contains the induced partition.
c. $\text{EXH}(A^*)$ is the induced partition.

When these conditions are met, A is said to be strongly stable.

The proof of the theorem is straightforward and we can sketch it here; a more detailed version is in Section A.1 in the appendix. (c) to (b) is immediate, and for the simple definition of exhaustification we are using here, (b) to (c) follows from the observation that dominated propositions are exhaustified to contradictions. The equivalence between (a) and (b)/(c) is proven by verifying that propositions that are the true answer somewhere are exhaustified to partition cells, and that φ is always the strongest true proposition wherever $\text{EXH}_A(\varphi)$ is true. The appendix also contains a proof that this result extends to the minimal world and the innocent exclusion operators.[24]

As an example of applying the theorem, our earlier A_0 (or A_1 for that matter) has no

[24]One subtlety of this result that is worth drawing attention to is the implication from (b) to (c). We can understand strong stability as the requirement that we should be able to express the induced partition in *some* way. In the formulation of (b), we tolerate that some alternatives might not be useful to express the partition, although they are still present as alternatives and active in exhaustification. The implication to (c) establishes that these alternatives are the dominated ones, and that we might as well completely ignore them: their presence or absence does not affect the exhaustification of other alternatives.

This approach to "extra" alternatives differs from that of Fox (2020). Fox offers a form of this result and argues, as we do here, that it is desirable for exhaustification to partition out logical space, but also finds differences between that condition and STA (which he argues is inferior). This does not contradict our result because he defends the innocent inclusion and exclusion definition of exhaustification, for which the result does not hold, and also because the requirement he proposes is that $\text{EXH}(A)$ should be exactly the induced partition and not merely contain it: thus there should be no "extra" alternatives. This requirement would be problematic for us, because we want to find (for instance) that disjunction is a "good" connective, even though $p \vee q$ is dominated when p and q are also available. Fox's argument has to do with free choice readings of disjunction and mention-some questions and is beyond the scope of this paper; note that mention-some questions are a category of questions that are known to violate the strongest answer condition. See also Section 4.3 for more discussion of the role of dominated alternatives.

single strongest true propositions when p and q are both true; thus it does not verify STA, and per the theorem, EXH(A_0) does not contain $\mathcal{P}(A_0)$, as one may also verify directly. Another reason A_0 fails to verify STA is that no proposition is true when both p and q are false. Some examples of extensions of A_0 that fix these two problems, and therefore that are strongly stable, are given in (27). A_2 is the most straightforward fix: it adds propositions corresponding to the two problematic partition cells; the other two cells can be obtained from exhaustifiying p and q. A_3 is exactly the set of upward-monotonic functions of p and q; note the presence of a tautology (whose exhaustification is $\neg p \wedge \neg q$) and the fact that $p \vee q$ and the contradiction $\bot$ are dominated. In A_4, which contains the induced partition, p and q are dominated and exhaustifying the non-dominated alternatives is vacuous. Assuming that p and q are expressible without logical operators and that all other propositions under consideration are more complex, A_4 being the alternative set is less efficient than A_2 or A_3: well-informed speakers will never use the simplest possible sentences.

(27) $A_2 = \{p, q, p \wedge q, \neg p \wedge \neg q\}$
$A_3 = \{p, q, p \vee q, p \wedge q, \top, \bot\}$
$A_4 = \{p, q, \neg p \wedge \neg q, \neg p \wedge q, p \wedge \neg q, p \wedge q\} = A_0 \cup \mathcal{P}(A_0)$

3.2 The consequences of stability on lexicalization choices

3.2.1 Background and assumptions

We have argued in the previous section that it is desirable for the alternative set A to be strongly stable, that is, to be such that exhaustification produces a partition of logical space, and that this is equivalent to verifying STA. In this section, we will see how a stability requirement affects lexicalization choices.

Here we assume that A is derived as the structural theory predicts: it contains all meanings expressible using sentences of a certain shape, up to a certain level of complexity. We will remain vague as to what kind of sentences are under consideration, and we will instead adopt the assumption that the meanings in A can be expressed as logical combinations of n logically "atomic" propositions $p_1, \ldots, p_n$. For the sake of concreteness, it could be that A contains all propositions expressible using a sentence of the form in (28a), with DET a one-word determiner and the domain of people being constant — here we see how the lexicon shapes A. A could also be formed of "simple" quantifier phrases in the sense of Barwise and Cooper (1981) over a given domain, as in (28b). In both of theses cases, at the semantic level, the elements of A are logical combinations of the "atomic" propositions expressed by the sentences in (28c). We will assume that all the atomic sentences are part of A.

(28) a. DET people came to the party.
b. QP came to the party.
c. Ann/Bill/Carol/... came to the party.

These assumptions incorporate certain simplifications relative to naturalistic alternative sets. In considering that all sentences are, at the semantic level, logical combinations of sentences without connectives, we ignore the possibility of modality or conditionals. We also set aside the asymmetry predicted by the structural theory in deriving alternatives: while complex sentences compete with simpler ones, simple sentences are not in general assumed to evoke more complex ones as alternatives. Here we assume that a single set A is used for the interpretation of every sentence. These simplifications are necessary to arrive at mathematical results at this stage — the exploration of more realistic models is left for future work.[25]

In our setting, there are 2^n possible situations, and 2^{2^n} possible propositions. Each proposition φ can be associated to a logical function f of arity n through the relation in (29). For instance, going back to (28a), if DET is *all*, f is the function that is true only when all its inputs are true. If φ is one of the atoms p_i, then f is the ith projection π_i, defined in (30). We will call F the set of logical functions from which the propositions in A are obtained. While the logical functions in F are not exactly the same objects as the logical operators of the lexicon, they are closely related. In particular, the functions are upward-monotonic whenever the operators are. Below, we are going to review a series of results establishing that only upward-monotonic functions should be expressible if we want A to be stable.

(29) $\varphi \equiv f(p_1, \dots, p_n)$

(30) $\pi_i(p_1, \dots, p_n) := p_i$

Given our discussion of indirect implicatures in Section 2.2, it is also interesting to look at what happens when elements of A are negated. Thus not just A itself should be strongly stable, but also $\tilde{A}$, defined in (31). If that is so, then A should verify the *weakest false answer* condition (WFA): at every world w, $\mathbb{F}_A(w)$ should have a maximal element for entailment. In some of our results below, we will need to appeal to both STA and WFA to single out upward-monotonic functions; this reflects our discussion of the problems with indirect implicatures in Section 2.2. Sets A that verify both STA and WFA will be called bilaterally strongly stable.

(31) $\tilde{A} := \{\neg\varphi : \varphi \in A\}$

In some cases, we will look not just at STA/WFA, but also at non-atomic STA/WFA, defined as follows: all true sets (resp. false sets) are such that either they have a minimal (resp. maximal) element, or all their minimal (resp. maximal) elements are atoms. Sets that verify non-atomic STA/WFA will be called (bilaterally) weakly stable. One reason for considering the property of weak stability is that it turns out to be equivalent to the property that Bar-Lev and Katzir (2022a) call "stability", as we will come back to in Section 3.3. Using this notion, we will be able to recapitulate their results. Weak stability can intuitively be understood as a requirement that the exhaustified meaning of utterances should not overlap with that of their alternatives. For Bar-Lev and Katzir (2022a), this property is desirable to reduce context-dependency in production (and therefore also in comprehension, assuming Bayesian listeners); see Section 3.3. One may also argue that it is desirable in the same sense strong stability is, by enforcing that exhaustification reinforces the meaning maximally.[26]

As a final note, it is not true in general that bilateral strong stability cannot hold of sets of propositions corresponding to non-upward-monotonic functions. If F comprises all functions that are true at exactly one of the 2^n possible inputs, and all functions that are false at exactly one of the 2^n possible inputs, then the corresponding A verifies both STA and

[25] See also Section 4.3 for more discussion of the limitations of our formal setting.

[26] Unlike strong stability which can only really be understood as a property of the entire set A, weak stability can thus be seen as a property of each individual utterance. In particular, the "global" requirement made by strong stability that the set of exhaustified propositions should cover logical space is absent in weak stability. If we see weak stability as a property that distributes down to utterances in this way, the fact that it places no requirement on atoms also makes sense: looking at utterances one by one, we should take into account that "being an alternative to" is not a symmetric relation, and that atoms do not necessarily compete with other utterances. I thank a reviewer for suggesting these conceptual underpinnings of weak stability.

WFA — we already discussed a set with such a shape in Section 2.2, and the set A_4 in (27) also examplifies this approach toward verifying STA. Thus we are going to have to add additional restrictions on what candidates are considered to make it so that only upward-monotonic functions are suitable; we will go though several ways of doing it. The extra conditions will be properties that simple natural language expressions verify in practice; we assume that they have independent sources and do not attempt to explain them here.

3.2.2 Two-place functions

In the case where $n = 2$, F can be seen as a set of binary connectives. There 16 possible connectives, of which 6 are not really binary (the two constants, the two projections, and the negations of the projections). Of the remaining 10, which we will call the genuinely binary connectives, 6 are symmetric (conjunction, disjunction, bi-implication, NAND, NOR, XOR) and 4 asymmetric (left- and right-implication and their negations).

The following results can be proven:

(32)
a. If A is strongly stable (resp. strongly stable under negation), conjunction (resp. disjunction) is part of F.
b. If A is bilaterally strongly stable, and the atoms are not dominated, then F contains exactly the upward-monotonic connectives: conjunction, disjunction, both projections, and both constants.
c. If all non-atomic elements of A correspond to symmetric connectives, and A is bilaterally weakly stable, then F does not contain any other genuinely binary connectives than conjunction and disjunction.

Proofs are omitted here since it is a matter of enumeration, and we are going to cover more general cases below.

Result (32a) establishes that upward-monotonic connectives are necessary to extend the atoms to a bilaterally strongly stable set, echoing our discussion of competition between atoms in Section 2.3. Result (32b) shows that adding non-upward-monotonic functions to the set F would either break stability, or lead to a system where the atoms could not be used informatively, which is inefficient since the atoms are simpler than other propositions.[27] Result (32c) shows that if we focus on symmetric connectives, the weaker requirement of bilateral weak stability is enough to single out the upward-monotonic connectives.[28]

3.2.3 Symmetric functions

Moving on to the general case where n is arbitrary, a natural requirement to place on F is that, other than the projections (corresponding to the atoms in A), it should only contain symmetric functions, that is, functions which are invariant to permuting the atoms, as per (33).[29] To take a concrete example, a sentence with the structure in (28a) where DET has been replaced with any one word-determiner of English (*some*, *few*, *three*, ...) expresses a proposition that is a symmetric function of the atoms. In order to express a non-symmetric

[27] As pointed out by a reviewer, in a model that incorporates the asymmetric nature of pragmatic competition, and where the atoms presumably do not have the more complex sentences as alternatives, it would be possible to use the atoms in informative ways even when they are dominated in A. For this reason, this particular result is only interesting if we accept that in at least some cases, basic sentences can have more complex sentences as alternatives; see also Section 4.3.

[28] This last result recapitulates the findings of Bar-Lev and Katzir (2022a); cf. Section 3.3.

[29] There is no relation between this use of the word "symmetry" and the notion of symmetric alternatives.

function, we need a subject as complex as *Everyone but John* or *The two youngest people*. The deeper reason this holds is that determiners and simple quantifiers express set-theoretic relations between predicates without reference to specific individuals: the only thing they can depend on is the cardinality of the various regions in the Venn diagram of their two predicate arguments. This property is identified as a universal of "logicality" by Keenan and Stavi (1986), and (in a slightly different form) of "isomorphism invariance" by Peters and Westerståhl (2006); Steinert-Threlkeld and Szymanik (2019) call it the "quantity universal".

(33) **Symmetric functions:** the Boolean function f is symmetric iff for any permutation σ of $\{1, \ldots, n\}$:

$$f(p_{\sigma(1)}, \ldots, p_{\sigma(n)}) \equiv f(p_1, \ldots, p_n)$$

The following results can be proven:

(34) a. For $n \geq 3$, if all functions in F other than the projections are symmetric, strong stability is impossible.

b. If all functions in F other than the projections are symmetric, and A is weakly stable, the only non-trivial non-atomic propositions that can be in A correspond the universal, existential, and negated existential quantifiers over the atoms, as well as the disjunction of the universal and the negated existential.

c. If all functions in F other than the projections are symmetric, and A is bilaterally weakly stable, the only non-trivial non-atomic propositions that can be in A correspond the universal and existential quantifiers over the atoms.

Detailed proofs are found in Appendix A. For (a) through (c), the general idea is that the strongest true proposition at a world where several atoms are true needs to entail all the true atoms, which is incompatible with the restriction to symmetric functions except in extreme cases.

Interestingly, if we apply (34b) to the case of quantifier phrases as in (28b), it singles out almost exactly the logical functions that can be expressed with a one-word quantifier in English: *somebody*, *everybody* and *nobody* — these are also the three corners of the Square of Opposition other than the O-corner. It also allows for a function paraphrasable as *nobody or everybody*, though the corresponding proposition is dominated and will never be used by an informed speaker. The point of presenting (34b) separately from (34c) is to show that the weaker constraint of simple weak stability actually gets us most of the way to the result.

3.2.4 Projectively symmetric functions

We have seen that symmetric functions as a class are too restricted to allow for strongly stable sets of alternatives. A larger domain of functions we can consider instead is projectively symmetric functions: these are obtained by composing a symmetric function with a projection to a subset of atoms, as defined in (35).

(35) **Projective symmetry**: f is projectively symmetric iff there is $i_1, \ldots, i_k$ a subset of $\{1, \ldots, n\}$ and g a symmetric function of k arguments such that:

$$f(p_1, \ldots, p_n) \equiv g(p_{i_1}, \ldots, p_{i_k})$$

Relative to the same atoms as in (28), sentences of the form given in (36), with one-word determiners and arbitrary restrictors, express projectively symmetric functions.[30] Another case where simple sentences express projectively symmetric functions is that of (28b) for simple quantifier phrases. Ultimately, in both cases, projective symmetry follows from the fact that natural language determiners are conservative (Barwise and Cooper 1981), together with the aforementioned logicality universal of Keenan and Stavi (1986).

(36) DET RESTRICTOR came to the party.

The following results can be proven:

(37) a. If F contains only projectively symmetric functions and A is strongly stable, then all conjunctions of atoms are in A, and all propositions in A that entail the disjunction of all atoms correspond to an upward-monotonic function in F.
b. If F contains only projectively symmetric functions and A is bilaterally strongly stable, then A contains all conjunctions and disjunctions of atoms and all functions in F are upward-monotonic.

The proof is detailed in Appendix A; again, the key idea is that propositions that are the strongest true answer somewhere need to entail certain atoms, which for a projectively symmetric function is only possible by being a conjunction of atoms; in turn all other expressible functions need to also be upward-monotonic to be entailed by the strongest true answers.

3.2.5 Convex functions

As an alternative to projective symmetry, another constraint on meanings that might be active is convexity. As already mentioned, there is evidence that non-convex functions are harder to reason about (Chemla et al. 2019) and some semantic and pragmatic phenomena can be explained by a constraint banning non-convexity (Enguehard and Chemla 2021; Solt and Waldon 2019). The examples we have been considering would all contain exclusively convex functions of the atoms, due to the absence of non-convex determiners.

While a convexity constraint does come close to singling out upward-monotonic functions, the result is not as clean as the previous ones:

(38) If all functions in F are convex, and A is bilaterally strongly stable, then define $\bar{A}$ in the following way:

$$\bar{A} := \{\varphi \vee \left(\bigwedge_i p_i\right) : \varphi \in A\}$$

$\bar{A}$ contains all conjunctions and non-empty disjunctions of atoms, and all of the propositions in $\bar{A}$ correspond to upward-monotonic functions.[31]

As before, the proof is detailed in Appendix A.

[30] One limitation here is that we are not modelling the use of logical words in restrictors and the potential structure in the set of alternatives that arises from that. See also Section 4.3.

[31] The way to make sense of this result is that A is almost a set of propositions corresponding to upward-monotonic functions, but it deviates from it only in that the propositions it contains may be false in the situation where all atoms are true. In other words, the functions in F are true at all inputs higher than other true inputs (like an upward-monotonic function), with the possible exception of the input that is all ones.

3.2.6 Closed classes of functions

Our final result takes a more structured perspective on the set of expressible functions. In the above, we have considered that any set of logical functions where each individual function satisfies some constraint we have placed is up for lexicalization. We also have considered each number of atoms separately. In natural language, of course, there is a finite number of logical words, and they mostly apply to domains of arbitrary arity. To model this, one approach we can adopt is to assume that the set of expressible logical functions is obtained through composition of a finite number of primitives.

We will assume that we have a set B of logical primitives, and that the set of expressible functions of any arity is obtained by composing these primitives with projections in arbitrary ways. Such a structure is called a clone, as per the definition in (39). For a given choice of B, the clone we are interested in is the smallest clone including B, which is said to be generated by B, and of which B is said to be the basis. In the case where B generates the set of all possible logical functions $\mathbf{L}$, it is said to be functionally complete: famously, $\{\wedge, \neg\}$ is functionally complete, as is {NAND}. The upward-monotonic functions form a clone $\mathbf{M}$, for which a possible basis is $\{\top, \bot, \wedge, \vee\}$.

(39) **Clone:** let F be a collection of logical functions of varying arities, and F_n be the elements of F of arity n. F is a clone if:

a. All the projections are in F: for any n and any $i \leq n$, the function π_n^i such that $\pi_n^i(p_1, \dots, p_n) = p_i$ is in F_n.

b. F is closed under composition: if $f_1, \dots, f_k$ are in F_n and g is in F_k then $h : \vec{p} \longmapsto g(f_1(\vec{p}), \dots, f_k(\vec{p}))$ is in F_n.

For a clone F, we get a set of logical functions F_n for any arity, to which we can associate a set of propositions A_n in the same way as before;[32] thus it makes sense to ask whether a clone is stable. The following results can be obtained:

(40) Assume the sets of expressible functions are obtained by taking the functions of the appropriate arity in a given clone F that is a proper subset of $\mathbf{L}$.

a. If A_n is bilaterally weakly stable for all n, then $F \subseteq \mathbf{M}$.

b. If A_n is bilaterally strongly stable for all n, then $F = \mathbf{M}$.

The proof is detailed in Appendix A. It is straightforward and uses a result due to Post (1941):[33] there are exactly five maximal proper subclones of $\mathbf{L}$, including $\mathbf{M}$. The other four are the truth-preserving functions $\mathbf{P}_1$, the falsity-preserving functions $\mathbf{P}_0$, the affine functions $\mathbf{A}$ and the self-dual functions $\mathbf{D}$[34] — notice that downward-monotonic functions do not constitute a clone, since they are not closed under composition. To prove our result,

[32]The sets of propositions A_n could also be derived directly through the following generative rules:

(i) a. For any i, p_i is a proposition in A_n for all $n \geq i$.

b. For any $f \in B$ of arity k, and any k propositions $\varphi_1, \dots, \varphi_k$ in A_n, $f(\varphi_1, \dots, \varphi_k)$ is a proposition in A_n.

[33]This result is presented on Wikipedia in a significantly different and more modern way than the original publication: `https://en.wikipedia.org/wiki/Post's_lattice` (accessed April 2023).

[34]The truth-preserving functions are the functions such that $f(\vec{1}) = 1$. The falsity-preserving functions are the functions such that $f(\vec{0}) = 0$. The affine functions are the functions such that $f(\vec{p}) = a_0 \oplus \bigoplus_{i=1}^{n} a_i \wedge p_i$ for some (a_i); a basis is $\{\oplus, 1\}$. The self-dual functions are the functions such that $f(\neg p_1, \dots, \neg p_n) = \neg f(p_1, \dots, p_n)$; a basis is $\{\neg, \text{maj}\}$ where maj is the majority judgement function of arity 3.

it is essentially sufficient to verify that none of these can have subsets verifying STA, other than **M**.

What this result shows is that while we can achieve stability if all logical functions are easily expressible, the only way to achieve them with a more restricted vocabulary is by focussing on upward-monotonic expressions. In other words, if there is any restriction to the set of easily expressible propositions, as one might ague that there should be for reasons of economy, then that restriction should be to the upward-monotonic functions. This echoes our discussion in Section 2.2.

3.3 Link to Bar-Lev and Katzir's (2022a) communicative stability

The results we have gone through in this section can be seen as a generalization of the proposal of Bar-Lev and Katzir (2022a), who derive the upward-monotonic bias from a property of sets of alternatives they call communicative stability. Bar-Lev and Katzir's stability is defined in the context of probabilistic models of pragmatics where speakers choose their productions by optimizing a certain utility function that captures a trade-off between informativity and cost. The calculation of utility involves a probability distribution over worlds, representing the epistemic priors of participants; thus probabilistic prior beliefs — by which we mean beliefs on what is more or less likely, as opposed to beliefs on what is possible or not — can affect the choice of messages in a given situation. According to Bar-Lev and Katzir, this is an undesirable property: to make interpretation more robust, alternative sets should be such that production is not affected by probabilistic beliefs; this last property is what they call communicative stability.

Bar-Lev and Katzir go through all possible lexica of symmetric binary connectives, and check whether the alternative set obtained from all possible ways of combining two atomic sentences and one connective, for a given lexicon, is stable in their sense. They find that in order to have both communicative stability, and communicative stability for the set of negated alternatives, we cannot lexicalize anything other than conjunction and disjunction. This is the same result we had in (32c), because Bar-Lev and Katzir's stability is in fact equivalent to our weak stability (under reasonable assumptions on the set of alternatives and the sort of probability distributions being considered); if we remove the special case around atoms from their definition, we get what we have been calling strong stability. A proof of these equivalences can be found in Section A.3 in the appendix. Thus our results generalize Bar-Lev and Katzir's (2022a) approach to arbitrary number of atoms and to the notion of strong stability, while offering a new motivation for the constraint grounded in considerations of implicature derivation rather than in probabilistic models of pragmatics.

4 Discussion

4.1 Putting it all together: motivating the upward-monotonicity bias through exhaustification

In Section 1 of this article, we have seen that the logical property of monotonicity, and especially upward-monotonicity, has some kind of distinguished position in natural language: in particular, the logical lexicon is made chiefly of upward-monotonic operators. We call this fact the upward-monotonic bias. We have defended a strong characterization of the upward-monotonic bias, according to which non-upward-monotonic operators are systematically more complex than similar upward-monotonic ones, and the simplest operators are all upward-monotonic.

In Section 2, we have seen that theories of certain phenomena at the semantics-pragmatics interface, such as implicature derivation or question semantics, are dependent on the existence of an asymmetry between upward-monotonic propositions and other propositions to make correct predictions. This asymmetry is predicted by Fox and Katzir's (2011) structural theory of alternatives, but only because of the upward-monotonic bias.

In Section 3, we have formalized the requirement that one should be able to derive informative implicatures as the condition that exhaustification of the set of alternatives should give us a maximally fine-grained partition of logical space. This requirement is in turn equivalent to Dayal's (1996) strongest true answer condition, formulated in the context of question semantics, and applied here to alternative sets. We have presented a series of results to the effect that, under various additional assumptions, if the alternatives are obtained from a certain set of easily expressible functions (as the structural theory predicts) applied to independent atoms, the only way to satisfy our requirements is to make sure those functions are all upward-monotonic. Thus, the requirement that exhaustification should produce a partition predicts the upward-monotonicity bias (in its strong formulation). This is a generalization of Bar-Lev and Katzir's (2022a) result that the preference for upward-monotonic binary connectives follows from a requirement of communicative stability, since this latter notion is in fact equivalent to some form of the strongest answer condition.

If one looks into the proofs, the deeper reason that all our results hold is that exhaustification into a partition obtains when the complex alternatives formed with logical words are in certain entailment relations to simple sentences. Now, in the absence of logical words, simple sentences express upward-monotonic functions of the predicates they contain.[35] More complex sentences have to be upward-monotonic / order-preserving as well to preserve the entailment relations induced by simple sentences.[36]

Taking a broader view, I believe this collection of facts can be seen to support a certain narrative when it comes to explaining the upward-monotonic bias. Natural language is presumably subject to certain pressures towards economy. Among other things, the size of the lexicon and number of conceptual primitives should be limited. This plausibly precludes the ability to express arbitrary logical functions at the semantic level, so that the logical lexicon has to be drawn from some restricted class, built compositionally out of a small number of primitives, in a way similar to the clones of Section 3.2.6. At the same time, there are also pressures towards communicative efficiency: speakers should be able to say what they want to say in reasonably brief ways. Ideally, pragmatic reasoning is able to bridge this gap, and get rid of the inaccuracy imposed on the semantics. However, not all possible languages

[35]This is because semantic composition operations are upward-monotonic: if we follow for instance Heim and Kratzer (1998), the semantics of simple extensional sentences can be derived in a system two basic compositional operations: functional application (FA), and predicate modification (PM). PM essentially perfoms conjunction, while FA applies a functor to its argument, which is an upward-monotonic operation with respect to the functor. Furthermore, as far as DPs are concerned, the simplest functors, lifted proper names and definite descriptions, are upward-monotonic.

[36]This is most clearly seen in relation to the minimal-world definition of exhaustification (Section A.1). For exhaustification not to be vacuous, there should be some non-trivial order on worlds; if all worlds are minimal, exhaustification will not do anything. The atoms induce a certain order on worlds already — a complete lattice isomorphic to the powerset of the set of atoms — and adding further propositions can only remove structure. In fact, adding any proposition that is not an upward-monotonic function of the atoms will remove some relations; keeping to upward-monotonic functions is the only way to preserve the initial lattice.

lend themselves equally well to pragmatic reasoning: if our alternative sets exhibit certain configurations, we will end up with symmetric alternatives, and pragmatic computations will yield trivial or degenerate results.

Upward-monotonicity is a solution to all these constraints. With an upward-monotonic logical lexicon, pragmatic reasoning avoids the symmetry problem, and we can reliably derive strong implicatures. At the same time, speakers can express themselves in simple ways: for instance, they can specify the denotation of a predicate by listing its members using a conjunction and the implication that the predicate is false of all unlisted individuals will go through. The logical lexicon is also simple: as described by Katzir and Singh (2013), a language of upward-monotonic operators can be seen as having just two primitives. What our results suggest is that upward-monotonicity is in fact the *only* solution: to get maximally strong implicatures, we need STA, as well as WFA for indirect implicatures, and in turn, to get STA and WFA, under plausible additional constraints on the lexicon, we need upward-monotonic operators only. The function of the upward-monotonic bias, then, is to avoid the symmetry problem and guarantee that pragmatic computations will perform their task of meaning enrichment, letting speakers achieve high degrees of precision while keeping the lexicon and the semantics at a tractable level of complexity.

4.2 What counts as an explanation for the upward-monotonic bias?

We have noted in the first section that the upward-monotonic bias, especially through its incarnation as the problem of the missing O-corner, has been the subject of a lot of literature. Of course, earlier authors have already put forward a number of explanations for the bias. The argument of this section is that our proposal explains the bias in a deeper sense that earlier ones.

As we have seen, to explain the asymmetry between the lexicalized I-corner (*or*, *some*...) and the missing O-corner (**nand*, **nall*...), Horn (1973) appeals to the notion of markedness: because negation is marked, negative operators are dispreferred. The problem with markedness as an explanation is that it is merely a restatement of the observation in more general terms: markedness, for Horn (1973), is diagnosed on the basis of morphological complexity, so it cannot really explain an observation that is itself about morphological complexity.

To make the markedness analysis more precise, various authors have proposed that the meaning of logical operators is represented at some level in a language whose primitives are such that upward-monotonic operators are simpler to express. In particular, Katzir and Singh (2013) propose semantic primitives *min* and *max* that generalize disjunction and conjunction respectively, while Uegaki (2022) proposes a model of the logical lexicon's informativeness and complexity involving a representational language based on first-order logic. Another proposal of this kind, from a fairly different conceptual perspective, is made by Incurvati and Sbardolini (2022) for the specific case of binary connectives; they relate the absence of **nand* to consideration of conversational dynamics. We have ourselves assumed that some proposal like that is correct, when we claimed that downward-monotonic operators are systematically more complex. However, we have presented it as a description rather than an explanation. Indeed, these proposals adopt a logical language where lexicalized operators are simpler *because they want to fit natural language*; the claim that the logical language predicts lexicalization patterns is merely a refinement of the observation that attested operators are what they are. To be sure, those proposals are interesting as models of the logical lexicon that can make precise predictions on for instance relative complexity, but that does not bear directly on the "why?" of the upward-monotonic bias.[37]

We can make this point concrete by considering the case of first-order logic, used as a logical language by Uegaki (2022): the primitives of first-order logic are conjunction, disjunction, existential and universal quantification, and negation. All these operations are upward-monotonic save for negation, so that downward-monotonic or non-monotonic expressions need to involve negation somewhere and will as a consequence be more complex. This is not a necessary property of languages as expressive as first-order logic, and one may very well adopt an equally expressive language based on downward-monotonic or non-monotonic connectives, such as NOR, XOR, or $\exists!$. The reason the connectives of first-order logic are mostly upward-monotonic is that it came across as natural to the people who chose them, precisely because of the upward-monotonic bias.

It is also important to note that the point being made here applies whether we are talking about morphological representations directly, or about cognitive representations.[38] Upward-monotonic quantifiers in particular are known to be easier to reason about than downward-monotonic ones (e.g. Just and Carpenter 1971; see also the discussion and references in Szymanik 2016, sec. 5.5) as well as non-monotonic ones (Chemla et al. 2019). It is tempting to claim that we observe upward-monotonic functions in the lexicon *because* those functions are easier to process for the mind. Nevertheless, it is still not clear that the lexicalization facts and the cognitive patterns are separate observations, as there are plausible causal pathways both ways: while it seems reasonable to assume that we do not lexicalize concepts that are hard to think about, it is also reasonable to think that concepts that are harder to express should be harder to reason about, because their mental representation is more complex — under the assumption that either natural language is used by its speakers to help their thought, or that there is a compositional language of thought that has structural similarities to natural language. In fact, models of learning or processing of quantifiers that are purely semantic, that is, that do not involve symbolic representations, do not predict a difference in difficulty between upward- and downward-monotonic quantifiers; this is true of semantic automata (e.g. Szymanik 2016) as well as of the neural network model of Steinert-Threlkeld and Szymanik (2019). As an aside, we can note that speakers do use downward-monotonic or non-monotonic operators quite commonly, in spite of the complexity involved in expressing them. In fact, due to scalar implicatures, statements involving monotonic operators are routinely interpreted as if non-monotonic operators had been used — for instance, *some* as *some but not all*. This seems hard to reconcile with the idea that non-monotonic or downward-monotonic operators are so challenging to process that lexicalization is impossible.

It follows from the previous discussion that we want to derive the upward-monotonic bias from some principle that is motivated without reference to the morphological observations we started from. One proposal that meets this challenge is that of Enguehard and Spector (2021), who derive the dispreference for O-corner operators from an independent generalization about the denotation of NPs, and considerations of probabilistic informativ-

[37]To be sure, Katzir and Singh (2013) do point out that their proposal has good consequences for implicature derivation in the *some/all* case and offer this fact as part of the motivation for it, prefiguring the explanation we have offered. Unlike us however, they do not show that it is *necessary* for the conceptual language to be this way.

[38]Here we can note that while we have been assuming that downward-monotonic operators are morphosyntactically complex, our line of explanation could also work if the characterization of the upward-monotonic bias that we had adopted was that they are complex at the level of cognitive representations. One would need to assume that alternatives are derived from semantic and not syntactic representations, as Buccola et al. (2022) propose, and that the structural theory applies to those semantic representations.

ity. However, the dispreference for the O-corner is just one particular instantiation of the upward-monotonic bias: Enguehard and Spector's (2021) proposal does not directly explain why downward-monotonic operators are dispreferred in general. They also only discuss potential lexica based on monotonic operators, and it is not clear that the general approach could be extended to non-monotonic lexica. Thus, while they can potentially explain the choice between upward- and downward-monotonicity in the lexicon, they cannot explain why we prefer monotonicity in general.

The approach of deriving the upward-monotonic bias from constraints on alternative sets, introduced by Bar-Lev and Katzir (2022a) and taken up here, avoids the problems we have just described, as it does not assume any particular representational language, and does not focus on a particular choice between lexica; instead we consider all possible choices of connectives within certain restrictions. The explanation we have provided is primarily dependent on theories of scalar implicatures. These theories can be derived or at least motivated through Grice's maxims, which are very broad principles of communication, and do not incorporate our observations about the logical lexicon in any way. In this sense, our explanation goes beyond a mere description or restatement of the initial observation.

If we look in detail at the results of Section 3.2, our derivation of the *necessary* character of an upward-monotonic lexicon was dependent on additional constraints on meaning: we assumed, in sequence, that determiners have to express symmetric, or projectively symmetric, or convex functions, or that expressible functions have to form a clone. Thus, this part of our explanation is dependent on those stipulative constraints.[39] While the consideration of clones is arguably principled, the other restrictions have a primarily observational basis: natural language connectives express symmetric functions in practice, etc. Thus, in light of our earlier discussion, the degree to which our results help truly explain the upward-monotonic bias depends on the degree to which the observations underlying our constraints can be considered independent from the ones establishing the upward-monotonic bias and can be hoped to receive a separate explanation. When it comes to symmetric and projectively symmetric functions in particular, our results can be seen as reducing the upward-monotonic bias to the logicality and conservativity universals. The logicality universal, in turn, has been claimed to provide a learnability advantage (Steinert-Threlkeld and Szymanik 2019), while the conservativity universal has been argued to follow from syntactic constraints (e.g. Fox 2002, fn. 7).

4.3 Some challenges

Before we conclude, let us discuss some potential challenges to the argument presented above.

Limits of the formal setting The relevance of our formal settings to lexicalization patterns is in some cases questionable, and it is important to admit that the specific choices made in Section 3 were determined by what results could be obtained. We have already discussed the stipulative nature of the restrictions on function domains, but there also a variety of more technical objections worthy of note.

[39]However, thanks to the adoption of strong stability over weak stability, some of our results concerning binary connectives do not depend on a restriction to symmetric connectives, and in this respect improve on Bar-Lev and Katzir's (2022a) approach.

A first remark is that our results crucially depend on the assumption that complex sentences compete with simple ones: what pushes towards upward-monotonicity for lexicalized operators is the need to maintain entailment relations with simple sentences. It is standard to assume that "φ or ψ" has φ and ψ among its alternatives, but not so much that "Some P Q" or even "Somebody Q" competes with "John Q"; most discussions of the implicatures of quantifying words assume that alternatives are only obtained through replacing logical words by other logical words. Without atoms in the picture, there would be no issue lexicalizing exclusively downward-monotonic determiners, for instance: they would be neatly ordered by entailment and exhaustifying them would lead to the same upper- and lower-bounded meanings as our actual determiners give us. Thus, for our models to be relevant to natural language, we have to accept that in at least some cases "logical" sentences compete with "non-logical" ones. This property of the general approach is already noted by Bar-Lev and Katzir (2022a, sec. 5.4). The structural theory does generally allow for atoms to compete with quantified statements; in Fox and Katzir's (2011) account, atomic alternatives can be obtained at least in some cases by replacing the determiner phrase with a proper name; Bar-Lev and Katzir (2022a) also suggest a derivation of atomic alternatives based on replacing variables.[40] Furthermore, the assumption that atoms compete with quantified statements is particularly plausible in the case of question-answer pairs. The answer sets for *who*-questions are generally assumed to contain both atomic answers (based on proper names or definite desciptions) and conjunctions thereof, and there exist also cases where *who* ranges over higher-order quantifiers, including some only expressible in relatively complex ways (Spector 2008; Elliott et al. 2022; Xiang 2021).[41]

Continuing, our models fall short of accounting for the diversity of constructions that are possible in natural language. Recall that we have justified our discussion of the case of n atoms as a model of the use of quantifying expressions. Natural language quantifiers operate on predicates, and in many cases what we observe is quantifying determiners with two predicate arguments, a restrictor and a scope. In Section 3, we have essentially identified the scope predicate P to the sequence of atomic sentences it can form ($P(x)$), and left it at that. Thus, we fail to represent the fact that the predicate may be complex, and itself feature logical words. Similarly, our approach to the restrictor, when we discuss projectively symmetric functions, is simply to assume that arbitrary restrictions are possible within an overall domain. Again, we ignore whatever internal logical structure the restrictor may have. In the same vein, we have focussed on competition in simple positive sentences and under negation, while ignoring various other environments where the derivation of implicatures can play out differently: Free Choice inferences under modals (Fox 2007), projection of implicatures under presuppositional attitudes (Spector and Sudo 2017), etc. We have also ignored downward-entailing environments other than negation, such as the restrictor of universals, the antecedent of conditionals, and the scope of desire predicates.[42] And on this topic, we have provided no explanation for the fact that universal determiners like *every* are downward-monotonic with respect to their restrictor (and *most* or *half* are

[40] I thank a reviewer for clarifying this point for me.

[41] If we want to maintain this assumption only in the case of coordination, we might instead argue that conjunction and disjunction will be lexicalized since they are the only possible binary connectives, and then *every* and *some* will be lexicalized as well because they are generalized conjunction and disjunction respectively; once we have *every* and *some*, other determiners have to be upward-monotonic as well.

[42] Note though that while they license NPIs, conditional antecedents and desire predicates are not in fact downward-monotonic on all analyses; see for instance von Fintel (1999) and Crnič (2011) for discussion. Another environment whose monotonicity is unclear is questions, since they license NPIs but are not related by entailment in an obvious way; see e.g. Schwarz (2017) and references therein.

non-monotonic), so that the upward-monotonic bias, as a generalization, does not extend to restrictors.[43]

Another issue we have set aside is the role of dominated alternatives. In our definition of stability, we merely requested that *some* alternatives could be used to express the induced partition, and allowed for other alternatives to exist in the set regardless of what their exhaustified meaning is; per the proof, these latter alternatives are the dominated ones. We expect a cooperative and well-informed speaker never to use dominated alternatives, since some alternative that is also true and more precise is always available. What, then, is the point of lexicalizing expressions that give us dominated alternatives, and do we not predict that they should be absent? Several things can be said: first, an alternative that is dominated in positive cases can be non-dominated when embedded in a downward-entailing context: for instance "φ or ψ" or "Somebody/anybody Q" are useful under negation. Second, it is in fact the case that these alternatives cannot be used by well-informed speakers when they compete with atoms: thus "φ or ψ" always triggers ignorance inferences — which, in some analyses, betrays the fact that disjunction is always exhaustified together with a silent modal (Meyer 2013) — and "Many people Q" is not perceived as a valid answer to "Who Q?". Since we have ignored intensional expressions as well as contexts where logical sentences do not have atomic competitors, it is normal that we should predict a lot of "useless" dominated alternatives.

All in all, we have made a model of a specfic situation of extensional sentences made of binary connectives or quantifying expressions competing with "atomic" alternatives, and shown that in this situation (generalized) conjunction and disjunction are necessary and the only useful logical words. Other alternatives can be lexicalized without interfering with exhaustification as long as they are upward-monotonic, but they will not be used in this particular situation. This is in the author's view a good first step towards establishing that the upward-monotonic bias is a way of satisfying some constraints on exhaustification. Of course, future work involving more elaborate models that account for the environments we ignored and for the cases where atomic alternatives are not present, as well as for the use of dominated alternatives, will be welcome; one can hope that it will not undermine the present conclusions.[44]

Plausibility of the primitives Another technical issue about our formal choices concerns the generative model. In the sort of languages that we are considering, where complex logical functions are derived from low-arity primitives, the representation of "quantity" quantifiers like *most*, *two* or *many* is very complex, and grows rapidly with domain size.[45] The actual cognitive representation of quantifiers is very probably not dependent on domain size in this way, which makes the relevance of such a language to the cognitive representation of quantification dubious. Note however that our results was not dependent on any assumption that the basis of the clone / the set of primitives of the generative model is finite or limited in arity: we might very well make all n-ary conjunctions primitives — as

[43] Since what drives our results is the requirement for a sentence expressing $DET(P)(Q)$ to be in an entailment relation with its atomic competitor expressing $Q(\mathbf{a})$, it could be that restrictors are not subject to the upward-monotonic bias because $DET(P)(Q)$ never competes with $P(\mathbf{a})$.

[44] We can mention in particular Bar-Lev and Katzir's (2022b) proposal to redefine stability in a way that non-perfectly-informed speakers can be included in the model; they find that the new constraint also points towards the desirability of upward-monotonic binary connectives.

[45] For instance, if we take $n = 5$ and assume that *most* means that at least 3 atomic propositions are true, *most* will be represented by a disjunction of 10 terms. For $n = 100$, the number of terms in the disjunction representing *most* has 29 digits.

Katzir and Singh (2013) essentially propose.

Competition in other lexical domains We have formulated a principle that exhaustification should turn sets of alternatives into partition of logical space, and claimed that this helps explain lexicalization patterns in the logical domain. A natural question is how this principle, in connection with the structural theory of alternatives, should affect other areas of the lexicon. If our explanation for the upward-monotonic bias is correct, we might expect to find lexicalization patterns that "avoid" situation of symmetries everywhere. However, this will not necessarily translate into a preference for upward-monotonicity — if this notion even applies to the domain — because the alternatives might not include simpler sentences.

In practice, it is difficult to find clear cases where the principle can be assessed. Attitude predicates have been argued to compete with each other in a few cases, in particular to explain the inference from use of *believe* that the speaker has doubts about the embedded proposition, through competition with *know* (see for instance Percus 2006 and Sauerland 2008). In general however, the complex semantics of non-purely-epistemic attitudes and their presuppositional nature makes it so that different attitudes are not logically related to each other or to their prejacent and there is no clear evidence of scalar implicatures arising from competition in this domain. For these reasons, we do not make clear predictions on how our principle constrains the lexicon of attitudes. Another potential case is simple content words: (41) triggers an inference that can be analyzed as a scalar implicature. Here, a situation of symmetry could arise if some of the modifiers available to replace *Slavic* overlapped without being included in one another. It does seem that this is not the case here, and in content domains in general, but given that competition among content words is clearly a lot more constrained than the simple type-based categories used in the structural theory predict, it is again difficult to draw conclusions.

(41) Mary read a book about SLAVIC linguistics.
$\rightsquigarrow$ Mary did not read a book about Romance linguistics.

Yet another, clearer case is that of gradable adjectives: here, what is observed is consistent with our postulated principle. Gradable adjectives express lower bounds, so that when several adjectives describe the same quantity, they are ordered by entailment: e.g. *huge* entails *big*. Adjectives expressing upper bounds on the same quantity exist in some cases, e.g. *short* or *tiny*, but they are marked: their distribution is more restricted and their use triggers more specific inferences (see for instance Sassoon 2010 and references therein). Notice how this reflects closely what is observed with quantifiers; in fact, Horn (1989) groups these observations together with those on the logical lexicon we have reviewed earlier, as instances of the general markedness of negation. While we have not proven here that things have to be this way, this situation allows for smooth exhaustification into strong upper-bounded meanings, under the assumption that the marked adjectives include a negative morpheme that cannot be deleted to generate alternatives.

The desirability of insentivity to priors While this is not the motivation we have offered for the desirability of (bilateral) stability, the earlier, technically equivalent proposal of Bar-Lev and Katzir (2022a) is based on the idea that speakers should be able to communicate reliably without agreeing on probabilistic priors. It is however not clear that this is desirable at all. When this condition is violated, speakers can use the fact that their choice of message is dependent on the prior to signal things about the prior. This seems to be useful, and it is in fact observed: Enguehard and Spector (2021) offer the example in (42), where the choice between *some* and *not all*, when both are true, seems to be determined by considerations of likelihood. In fact, this very pattern of sensitivity to priors is essential to Enguehard and Spector's proposal of an explanation of (one specific reflex of) the upward-monotonic bias. At the same time, Fox and Katzir (2020) show that sensitivity to priors is not observed in many cases where models inspired by Bayesianism predict that it should be.

(42) Context: *at an international conference, most talks are expected to be in English.*
  a. Some talks at this conference are in French / ?English.
  b. Not all talks at this conference are in ?French / English.

(adapted from Enguehard and Spector 2021)

Without resolving the whole controversy here, it seems that some sensitivity to priors is observed, but in a constrained way relative to what probablisitic models often predict. One possible way of reconciling the pattern noted by Enguehard and Spector (2021) with our discussion is to assume that the prior only plays a role in determining the choice between the unembedded alternatives and the negated alternatives, while within each set of alternatives, STA is verified, so that the choice of the specific message used is not sensitive to the prior. This would be consistent with examples like (42) and compatible with the general approach of stability as an explanation for the upward-monotonic bias. In such a system, the speaker-oriented definition of stability that Bar-Lev and Katzir (2022a) use is not verified: production does depend on probabilistic priors. However, a certain listener-oriented version of stability would hold: there is no way for the listener's assumptions about the prior to lead them to an incorrect interpretation. If the listener hears a negated alternative, they know the speaker takes the prior to be in such and such way, and because we have stability within the negated alternatives, they know which set of worlds the speaker thinks we are in. I leave the exploration of the constraints set by such a notion of stability on the lexicon to future work.[46]

4.4 Conclusion

With the caveats discussed above, we have argued in this paper that a desideratum of generating strong implicatures predicts the dominance of upward-monotonic logical functions in language. This extends the independently motivated, but technically equivalent proposal of Bar-Lev and Katzir (2022a) for binary connectives. The common idea is that the upward-monotonic bias is caused by constraints on the logical relations that should hold between the alternatives involved in pragmatic processes. Our final claim in this paper is that the necessity of deriving strong implicatures, so as to let speakers be maximally precise, without bloating the lexicon, makes the upward-monotonic bias necessary.

This picture is painted in broad strokes, and the intention is not to suggest that the problem is solved. As we have discussed, our formal setting only roughly approximates the linguistic system, and it might be that with more sophisticated operationalizations of the question, our conclusions will be undermined in some respects. What we can note is that the explanation we offer here is not incompatible with other explanations that involve ad-

[46]The later proposal of Bar-Lev and Katzir (2022b), also focussing on the case of binary connectives, incorporates this idea.

ditional factors, such as a probabilistic notion of informativity (as Enguehard and Spector (2021) propose). It is also compatible with specific proposals about the conceptual primitives of the logical lexicon (e.g. Katzir and Singh 2013). At any rate, it appears that 50 years after Horn 1973, the relation and potential causal links between lexicalization universals and competitive accounts of pragmatics remain a fertile ground for research.

A Appendix: proofs

A.1 STA and partition through exhaustification

Proof of the main theorem (26) Recall that A is a set of alternatives of which A^* is the non-dominated subset as defined in (25). $\mathrm{NW}_A(\varphi)$ is the set of non-weaker alternatives to φ, as per (19). The equivalence relation $\sim_A$ between worlds holds when two worlds have the same true sets, as defined in Footnote 22.

We omit the straightforward proof of the following lemma: for any proposition φ, $\textsc{exh}_A(\varphi) \equiv \textsc{exh}_{A^*}(\varphi)$.

Suppose A verifies STA. Let φ be a non-dominated proposition, and take w a world where no strictly stronger proposition is true. There is a strongest true proposition at w; it cannot be strictly stronger than φ by assumption, so it must be φ. All propositions weaker than φ are true at w (since φ is), while all non-weaker propositions are false at w (by definition of φ). It follows, first, that $\textsc{exh}_A(\varphi)$ is true at w, and second, that $\mathbb{T}_A(w) = \{\psi : \varphi \vDash \psi\}$ (call this relation (*)). Let w' be another world where the strongest true proposition is ψ. By (*), if $\psi = \varphi$, then $w \sim w'$ (and the reverse implication also holds). If φ asymmetrically entails ψ, then φ is false at w' (by definition of ψ) and so is $\textsc{exh}_A(\varphi)$ Finally, if $\psi \in \mathrm{NW}_A(w)$, then $\textsc{exh}_A(\varphi)$ is false at w'. Thus $\textsc{exh}_A(\varphi)$ is true at w' if and only if $w' \sim w$. All this also applies to $\textsc{exh}_{A^*}(\varphi)$ by the lemma. The conclusion is that every non-dominated proposition is exhaustified into an induced partition cell, corresponding to the worlds where it is the strongest true proposition. By (*), different cells have different strongest true propositions, which implies that all elements of $\textsc{exh}(A^*)$ are disjoint. Finally, since A verifies STA, a strongest true proposition can be found at any world (and it will necessarily be non-dominated). We can conclude that $\textsc{exh}(A^*)$ is the induced partiton. This proves (a) to (c).

Suppose now that $\textsc{exh}(A^*)$ contains a partition. Let w be a world and φ the non-dominated proposition such that $\textsc{exh}_{A^*}(\varphi)$ is true at w. Due to the lemma, $\textsc{exh}_A(\varphi)$ is true at w. Since all elements of $\mathrm{NW}_A(\varphi)$ are false at w, φ is the strongest true proposition at w. Since $\textsc{exh}(A^*)$ contains a partition, such a φ can be found for any world. Generalizing to all worlds, A verifies STA. This proves (c) to (a) — in fact it proves something a bit stronger, and putting it together with (a) to (c), we get that if $\textsc{exh}(A^*)$ contains a partition, $\textsc{exh}(A^*)$ is the induced partition.

(c) to (b) is immediate, and (b) to (c) follows from the fact that dominated propositions are exhaustified into contradictions: if $\textsc{exh}(A)$ contains a partition, then that partition is in fact in $\textsc{exh}(A^*)$. By the above result, $\textsc{exh}(A^*)$ must then be the induced partition.

Extension to other exhaustification operators Here we extend (26) to more sophisticated definitions of the exhaustification operator; specifically, we are going to consider the "Minimal World" operator (MW, Schulz and van Rooij 2006; Spector 2006), and the "Innocent Exclusion" operator (IE, Fox 2007). The "Innocent Exclusion and Inclusion" operator (IE-II, Bar-Lev and Fox 2020), for which our result is false, will also be mentioned. Intuitively, the reason for these results is that MW and IE were mostly designed to behave in particular ways on dominated alternatives (for which the basic operator defined in (19) always outputs contradictions), and their behavior on non-dominated alternatives is roughly the same as the basic operator. In contrast, IE-II can do interesting things with dominated alternatives. For clarity, we will refer to the operator defined in (19) as $\textsc{exh}^0$.

The IE operator is defined in several steps:

(43) **Excludable set:** a set of propositions $S \subseteq A$ is said to be *excludable* in conjunction with a proposition φ if the intersection of φ with the negation of all members of S is not empty, or formally, if:

$$\varphi \cap \bigcap_{\psi \in S}(W - \psi) \neq \varnothing$$

We write $\mathrm{E}_A(\varphi)$ for the set of all subsets of A that are excludable in conjunction with φ.

(44) **Maximal excludable sets:** the maximal excludable sets $\mathrm{ME}_A(\varphi)$ of a proposition φ are the maximal elements of $\mathrm{E}_A(\varphi)$ under set inclusion.

(45) **Innocent exclusion:** a proposition ψ is *innocently excludable* in A relative to a proposition φ if it is an element of all elements of $\mathrm{ME}_A(\varphi)$. We write $\mathrm{IE}_A(\varphi)$ for the set of all innocently excludable propositions relative to φ in A.

(46) **IE exhaustification:** the IE exhaustification of φ in A is the conjunction of A with the negation of all its innocently excludable alternatives, or formally, the proposition $\textsc{exh}_A^{IE}(\varphi)$ defined by:

$$\textsc{exh}_A^{IE}(\varphi) = \varphi \cap \bigcap_{\psi \in \mathrm{IE}_A(\varphi)}(W - \psi)$$

The MW operator requires on order on worlds: such an order can be derived from the true sets, and is called the induced order.[47] The induced order is actually a pre-order: there are equivalence classes of elements that are related both ways. These equivalence classes are exactly the cells of the induced partition.

(47) **Induced order:** for a set of worlds W and a set of propositions A over W, the order over worlds induced by A, written as $\leq_A$ is defined as follows:

$$w \leq_A w' \iff \mathbb{T}_A(w) \subseteq \mathbb{T}_A(w')$$

(48) **MW exhaustification:** the minimal-world exhaustification of φ in A is the proposition obtained by restricting the denotation of φ to its minimal elements under $\leq_A$:

$$\textsc{exh}_A^{MW}(\varphi) = \{w : \varphi(w) \wedge \forall w'. [\varphi(w') \wedge w' \leq_A w] \rightarrow w \leq_A w'\}$$

A result due to Spector (2016) which we are going to make use of is that the IE alternatives can be defined in terms of the MW exhaustification:

(49) **Relation between IE and MW:** for any $\varphi \in A$:

$$\mathrm{IE}(\varphi) = \{\psi \in A : \textsc{exh}^{MW}(\varphi) \vDash \neg\psi\} \qquad \text{(Spector 2016)}$$

It follows in particular from (49) that $\textsc{exh}^{MW}(\varphi)$ is always at least as strong as $\textsc{exh}^{IE}(\varphi)$. Spector (2016) also shows that when the basic operator $\textsc{exh}^0$ does not yield a contradiction, the other two operators give the same result:

[47] Once we have an order on worlds, it makes sense to speak of monotonic propositions. In fact, all propositions in A are upward-monotonic with respect to the order induced by A. This is not necessarily very informative, because the induced order might be trivial.

(50) **Relation between simple and complex operators:** for any φ in A, if $\text{EXH}^0(\varphi)$ is satisfiable, then:

$$\text{EXH}^{IE}(\varphi) = \text{EXH}^{MW}(\varphi) = \text{EXH}^0(\varphi) \qquad \text{(Spector 2016)}$$

We will need to assume that A is finite. This guarantees that any proposition within a set entails some maximal proposition and is entailed by a minimal one. Because the induced partition is also finite, we can safely assume that all non-contradictory propositions have minimal worlds and that any non-minimal world is above some minimal world.[48]

We can now go through the proof. We will use (b-IE) to refer to clause (b) of (26) when EXH is defined as EXH^{IE}, and similarly for (b-0), (b-MW), (c-0), (c-IE) and (c-MW).

From Spector's result (50), it follows immediately that (b-0) entails (b-IE) and (b-MW), and (c-0) entails (c-IE) and (c-MW). Since (a) entails (b-0) and (c-0), (a) entails (b-IE), (b-MW), (c-IE) and (c-MW).

Suppose that $\text{EXH}^{IE}(A^*)$ contains the induced partition. Let w be a world and φ be a non-dominated proposition such that $\text{EXH}^{IE}(\varphi)$ is true exactly in the induced cell of w. Since φ has minimal worlds (here we use the fact that A is finite), $\text{EXH}^{MW}(\varphi)$ is non-empty; it must then be at least as big as an induced partition cell. Since $\text{EXH}^{MW}(\varphi)$ is at most as strong as $\text{EXH}^{IE}(\varphi)$, which is a partition cell, then $\text{EXH}^{MW}(\varphi) = \text{EXH}^{IE}(\varphi)$. Generalizing to all worlds, $\text{EXH}^{MW}(A^*)$ contains the induced partition (it also follows from the proof that (c-IE) entails (c-MW)).

Suppose that $\text{EXH}^{MW}(A^*)$ contains the induced partition. Let w be a world and φ a non-dominated proposition such that $\text{EXH}^{MW}(\varphi)$ is true exactly in the induced cell of w. If φ is true at some world w', there must be a minimal φ-world w'' somewhere below w' (here again we use the fact that A is finite). $\text{EXH}^{MW}(\varphi)$ is true at w'', so w'' is in the cell of w. Thus $w' \geq w$. Since φ is true anywhere above w by definition of the order, we can conclude that φ is true at w' if and only if $w' \geq w$. Now, for any ψ in A, if ψ is true at w, ψ is true anywhere above w, and therefore $\varphi \vDash \psi$. Thus ψ is the strongest true proposition at w. Such a φ exists for all worlds, so A verifies STA. In conclusion, if $\text{EXH}^{IE}(A^*)$ contains the induced partition, so does $\text{EXH}^{MW}(A^*)$, and the latter fact entails (a). In particular, this entails that (c-IE) and (c-MW) entail (a), and that if $\text{EXH}^{IE}(A^*)$ contains the induced partition, it is the induced partition (and the same for EXH^{MW}).

If φ is dominated, it can be verified that $\text{EXH}^{IE}(\varphi) \equiv \varphi$, and that φ cannot be true exactly at an induced partition cell. From this it follows that if $\text{EXH}^{IE}(A)$ contains the induced partition, $\text{EXH}^{IE}(A^*)$ contains the induced partition, and together with earlier implications this shows that (b-IE) entails (c-IE). It can also be verified that all minimal φ-worlds are also minimal worlds for some stronger propositions. From this it follows that if $\text{EXH}^{MW}(A)$ contains the induced partition, $\text{EXH}^{MW}(A^*)$ contains the induced partition, and together with earlier implications this shows that (b-MW) entails (c-MW). This completes the proof that (a), (b-MW), (b-IE), (c-MW) and (c-IE) are equivalent.

Some counterexamples showing the necessity of our assumptions From the above proof of (26) in the EXH^0 case, we can see that a slightly stronger result than what we presented is true: STA is equivalent to the fact that $\text{EXH}^0(A^*)$ is *any* partition, that partition necessarily being the induced partition. This fact does not extend to the IE and MW operators, as they can produce partitions that are coarser than the induced one in some cases.

[48] In Spector's result (49), the inclusion from left to right does not depend on the existence of minimal worlds, but I believe the inclusion from right to left does. See also the infinite case in the list of counterexamples below. The result in (50) does not depend on such an assumption.

Here is a case with six worlds where $\text{EXH}^{MW}(A)$ is a coarse partition, but $\text{EXH}^{IE}(A)$ has overlapping propositions: take $s = \{w_1, w_2, w_3\}$, $s' = \{w'_1, w'_2, w'_3\}$, $p = \{w_2, w_3, w'_3\}$, and $p' = \{w'_2, w'_3, w_3\}$. The induced order has two incomparable chains $w_1 \leq w_2 \leq w_3$ and $w'_1 \leq w'_2 \leq w'_3$; the induced partition distinguishes all six worlds. We have $\text{EXH}^{MW}(s) = \{w_1\}$, $\text{EXH}^{MW}(s') = \{w'_1\}$, $\text{EXH}^{MW}(p) = \{w_2, w'_3\}$, and $\text{EXH}^{MW}(p') = \{w'_2, w_3\}$; this is a coarse partition, with four cells. In this setting, $\text{EXH}^{IE}(p) = p = \{w_2, w_3, w'_3\}$ and $\text{EXH}^{IE}(p') = p' = \{w'_2, w'_3, w_3\}$, so that they overlap.

Here is a case with five worlds where $\text{EXH}^{IE}(A)$ is a coarse partition, and $\text{EXH}^{MW}(A)$ fails to cover the set of possible worlds. Take $s_1 = \{w_1, w'_1, w''\}$, $s_2 = \{w_2, w'_2, w''\}$, and $p = \{w'_1, w'_2, w''\}$. The induced partition distinguishes all five worlds. The induced order has two chains $w_1 \leq w'_1 \leq w''$, and $w_2 \leq w'_2 \leq w''$. We have $\text{EXH}^{IE}(s_1) = \{w_1\}$, $\text{EXH}^{IE}(s_2) = \{w_2\}$ and $\text{EXH}^{IE}(p) = p = \{w'_1, w'_2, w''\}$ so that $\text{EXH}^{IE}(A)$ is a 3-cell partition. However, $\text{EXH}^{MW}(p) = \{w'_1, w'_2\}$, so that no MW-exhaustified proposition is true at w''.

The assumption of finiteness is also important: without it, we can get a partition from the exhaustification of a set of alternatives that does not verify STA. Consider the following example: $W = \{w_i : i \in \mathbb{N}\} \cup \{v_1, v_2\}$, $s_n = \{w_i : i \leq n\}$ for all $n \geq 0$, $p = \{v_1, v_2\}$, $q = \{v_1\} \cup \{w_i : i\}$. The induced order has two incomparable chains, the infinite chain of the w_i's, and the finite chain of the v_i's, both ordered by decreasing index. There is no strongest true proposition in v_1. Nevertheless, EXH^{MW} does give us the induced partition: s_n is mapped to w_n for all n, p to v_2, ad q to v_1, thanks to the fact that q has no minimal worlds within the infinite chain. In this setting, p is excludable in conjunction with q and part of any finite excludable set, but it is not part of any maximal excludable set; the only such set is $\{s_n : n\}$. If we apply the definition of EXH^{IE} regardless of this oddity, we get again the induced partition.[49]

Finally, our result cannot be extended to the IE-II operator of Bar-Lev and Fox (2020). This is how Fox (2020), who adopts a simplified version of the IE-II operator, can argue that a partition through exhaustification constraint is superior to the strongest answer condition, even though we claim they are equivalent. The simplified IE-II operator used by Fox (2020) is given in (51). Like the IE operator, it denies all IE alternatives, but it also asserts all the non-IE ones. Thus, it always outputs either a contradiction, or a single partition cell.

(51) **IE-II exhaustification:** the IE-II exhaustification of φ in A is given by:

$$\text{EXH}_A^{IE-II}(\varphi) = \{w : \mathbb{F}_A(w) = \text{IE}_A(\varphi)\}$$
$$= \left(\bigcap\nolimits_{\psi \notin \text{IE}(\varphi)} \psi\right) \cap \left(\bigcap\nolimits_{\psi \in \text{IE}(\varphi)} (W - \psi)\right)$$

The crucial case for Fox's argumentation (in his Section 3) has the following structure: $p_1 = \{w_1, w'\}$, $p_2 = \{w_2, w'\}$, $q = p_1 \vee p_2 = \{w_1, w_2, w'\}$. Notice that q is dominated. Furthermore, there is no strongest true proposition at w'. Yet, because $\text{IE}(q) = \varnothing$, we have $\text{EXH}^{IE-II}(q) = q \wedge p_1 \wedge p_2 = \{w'\}$, so that $\text{EXH}^{IE-II}(A)$ is the induced partition.

A.2 Proofs of the results in Section 3.2

Recall that we consider both a set of propositions A, and a set of n-place logical functions F, linked together by the following relation, where the p_i's are n logically independent "atoms".

[49] I suspect some constraint on the shape of alternative sets can be found which is sufficient to prove the desired implication all while being generally satisfied in concrete scholarly analyses.

(52) $\varphi \equiv f(p_1, \ldots, p_n)$.

In this setting, the induced partition is that induced by the atoms: there are 2^n cells which are in bijective relation with valuations of the vector $p_1 \ldots p_n$. To simplify, we identify cells to indivual worlds. For any some subset I of $\{1, \ldots, n\}$, we will denote as w_I the world where the atoms whose index is in I are true, and all other atoms are false. There is a natural order on worlds: $w_I \leq w_J$ if and only if $I \subseteq J$. The same order can also be seen as the natural Cartesian order over bit vectors, or as the order induced by the atoms. We can then speak not just of a monotonic function but also of a "monotonic proposition", etc. Saying that a proposition φ is for instance upward-monotonic is equivalent to saying that it can be stated as $f(\vec{p})$ where $\vec{p}$ is the vector of the atoms' truth values and f is an upward-monotonic function.

For a set I and world w_I, we define two propositions γ_I and δ_I that will often be useful, as per (53). Note that γ_I is true at w_I and at any w_J with $I \subseteq J$, and false everywhere else, while δ_I is false at w_I and at any w_J with $J \subseteq I$, and true everywhere else. The choice of Greek letters is a mnemonic for "conjunction" and "disjunction" respectively.

(53) $\gamma_I := \bigwedge_{i \in I} p_i$
$\delta_I := \bigvee_{i \notin I} p_i$

Result (34) on symmetric functions Note that a symmetric Boolean function can only "count" the number of true atoms. Therefore, symmetric functions are in bijective relation to Boolean functions over $\{0, \ldots, n\}$.

We start by assuming only that A contains the atoms and verifies non-atomic STA and showing that this rules out any functions that have non-trivial truth conditions "in the middle" of the logical space. Consider a function f corresponding to a non-atomic proposition φ_f in A and let $K \subseteq \{0, \ldots, n\}$ be the set of integers k such that f is true when exactly k inputs are true. Assume that there is $k \in K$ and $k' \notin K$ such that $1 \leq k \leq n-1$ and $1 \leq k' \leq n$. Then, φ_f is not entailed by, nor entails, any atom. Let I be a subset of $\{1, \ldots, n\}$ of size k. φ_f is true at w_I, which means that the minimal elements of $\mathbb{T}_A(w_I)$ include a proposition ψ which is either φ_f or something stronger. Because φ_f is independent from the atoms, ψ is not an atom, and because A verifies non-atomic STA, φ' must then be the only minimal element of $\mathbb{T}_A(w_I)$. Then, φ' is stronger than all the p_i's for $i \in I$. This means that φ' is true at w_I, but not at w_J for any $J \neq I$ of size k (such a J exists because $k \leq n-1$). This is impossible, as φ' should correspond to a symmetric function. In conclusion, such k, k' do not exist: either $K \supseteq \{1, \ldots, n\}$, that is, f is a constant true function or an existential over atoms, or $K \subseteq \{0, n\}$, that is, f is a constant false function, a universal over atoms (roughly *everyone*), a negative universal (roughly *nobody*), or the disjunction of the previous two (*everyone or nobody*). This proves (b).

If A is strongly stable, it is also weakly stable and the previous result applies. Suppose $n \geq 3$, and take $I = \{1, 2\}$. Neither p_1 nor p_2 is the strongest true proposition at w_I since they are independent. The other atoms are false at w_I. By the above result, all complex propositions are false or weaker than p_1 and p_2 at w_I (remark that w_I is not the maximal world because $n \geq 3$). In conclusion, there is no strongest true proposition at w_I, contradicting our assumptions. This proves (a).

Suppose now that A is bilaterally weakly stable.. Consider $I = \{2, \ldots, n\}$. At w_I, p_1 is false, so by non-atomic WFA, all complex propositions must be either true at w_I, or either weaker or stronger than p_1. Within the set that non-atomic STA allows, this rules out the negative universal (*nobody*) and the disjunctive proposition (*everybody or nobody*). One may verify that if A contains all remaining propositions, non-atomic STA and WFA are verified. This proves (c).

Result (37) on projectively symmetric functions As a lemma, let us first prove that a projectively symmetric function non-trivially entailing an atom is a conjunction of atoms. Let f be a projectively symmetric function corresponding to a proposition φ_f such that $\varphi_f \vDash p_i$ for some i. We can find $J \subseteq \{1, \dots, n\}$ (the domain restriction of f) and $K \subseteq \{0, \dots, |J|\}$ (the accepted counts of f) such that for arbitrary I, φ_f is true at w_I if and only if $|I \cap J| \in K$. Suppose there is $k \in K$ such that $k \leq |J| - 1$. We can find $I_0 \subseteq J$ such that $i \notin I_0$ and $|I_0| = k$. Then, φ_f is true at w_{I_0}, but that is impossible because p_i is false at w_{I_0}. Hence no such k exists and either $K = \{|J|\}$, which means that φ is equivalent to γ_J (cf. (53)), or $K = \varnothing$, and φ is a contradiction. Note that in the former case, necessarily $i \in J$.

Taking the negation (noting that projective symmetry is closed under negation) and replacing atoms by their negation in the proof, we also obtain the following result: if f is projectively symmetric and $p_i \vDash \varphi_f$, then φ_f is either δ_J for some J containing i, or a tautology.

Now, for non-empty I, suppose there is a strongest proposition in $\mathbb{T}_A(w_I)$ and call it φ^*. Since φ^* is projectively symmetric, and it is entailed by p_i for all $i \in I$, φ^* is equivalent to γ_J for some J such that $I \subseteq J$ by the above lemma; furthermore since φ^* is true at w_I, J must in fact be I, and φ^* is in fact γ_I.

If A verifies STA, there is a strongest proposition at w_I for all I. By the above, that proposition is γ_I as long as $I \neq \varnothing$. Thus A includes all γ_I's for non-empty I. This proves the first part of (a).

Take now f in F such that φ_f is false at $w_\varnothing$, or equivalently, f is false for the all-zeroes input. Define $\mathcal{I}_f := \{I : w_I \in \varphi_f\}$, and $\psi_f = \bigvee_{I \in \mathcal{I}_f} \gamma_I$. Note that ψ_f, as a disjunction of conjunction of atoms, is upward-monotonic. For any I in $\mathcal{I}_f$, γ_I is a disjunct of ψ_f so that ψ_f is true at w_I. From this it follows that $\varphi_f \vDash \psi_f$. Furthermore, by the above, the strongest proposition in $\mathbb{T}_A(w_I)$ is γ_I, which means that $\gamma_I \vDash \varphi_f$, and in the end that $\psi_f \vDash \varphi_f$. In conclusion, ψ_f and φ_f are equivalent and φ_f is upward-monotonic. This proves the rest of (a).

If A verifies STA and WFA, the earlier results, applied under negation, tell us that the weakest false proposition at w_I with $I \subsetneq \{1, \dots, n\}$ is δ_I (which must then be part of A). It follows that any proposition false at w_I is at least as strong as δ_I, and therefore, is false at $w_\varnothing$. Then, any proposition true at $w_\varnothing$ is a tautology, and is therefore upward-monotonic. This proves (b).

Result (38) on convex functions We first prove that the weakest false answers at worlds other than the highest one are disjunctions of atoms. Define $N := \{1, \dots, n\}$. Let φ^* be the weakest false answer at w_I from some $I \subsetneq N$. φ^* is entailed by p_i for all $i \notin I$, so it is entailed by δ_I. In particular, φ^* is true at w_N. Suppose there is $J \subsetneq I$ such that φ^* is true at w_J. Because $w_J \leq w_I \leq w_N$ and φ^* is convex, φ^* must be true at w_I, contradicting our earlier assumptions. Hence there is no such J, which means that φ^* is exactly δ_I. All the non-empty disjunctions of atoms are in A and therefore also in $\bar{A}$.

Let φ^* be the strongest true answer at w_I for some I. φ^* entails all the p_i's for $i \in I$ and is therefore at least as strong as γ_I. If $I = N$, φ^* is necessarily γ_N (which is true at only one world). If $I \subsetneq N$, take J such that $I \subseteq J \subsetneq N$. If φ^* is false at w_J, it must entail the weakest false proposition at w_J, δ_J. This is impossible because δ_J is false at w_I, where φ^* is true. Hence φ^* is true at any such J. From this it follows that $\varphi^* \vee \gamma_N$ is equivalent to γ_I, and in general that all conjunctions of atoms are in $\bar{A}$.

To prove that all remaining propositions in A are mapped to upward-monotonic propositions in $\bar{A}$, one may verify that propositions in $\bar{A}$ are equivalent to a disjunction of conjunction of atoms constructed in the same way as in the previous proof.

Result (40) on closed classes of functions As said in the main text, there are five maximal proper subclones of the set of the all logical functions **L**, of which one is the upward-monotonic functions **M**, generated by $\{\top, \bot, \wedge, \vee\}$. If F is a proper subset of **L**, it is a subset of one of these five.

Two maximal proper subclones are the affine functions **A** and the self-dual functions **D**. The affine functions are affine combinations of the atoms, of the form $f(\vec{p}) \equiv a_0 \oplus a_1 \wedge p_1 \oplus \cdots \oplus a_n \wedge p_n$. The self-dual functions verify the property $f(\neg\vec{p}) \equiv \neg f(\vec{p})$. The affine and self-dual functions have in common that they are always true for exactly half of their possible inputs (except for the two constant functions in A).

Suppose non-atomic STA is verified for all A_n, and F is a subset of **A** or **D**. Let f be a non-constant element of F_n for some $n \geq 2$, and let I be a non-empty subset of $\{1, \ldots, n\}$ such that $f(\vec{p})$ is true at w_I (such a I exists since f is true for half its inputs). If $\mathbb{T}(w_I)$ has a minimal element corresponding to the function g, then there is at least one projection π_i such that $g \leq \pi_i$. Hence g cannot be the true constant, and it also cannot be the false constant since it is true at w_I. g is therefore true for half of its inputs, and since this is also the case of π_i, g is in fact π_i. Hence $f \geq \pi_i$, and because f is true for half of its inputs, $f = \pi_i$. If $\mathbb{T}(w_I)$ does not have a minimal element, all its minimal elements are atoms. There is at least one such element p_i that is stronger than $f(\vec{p})$, which means that $f \geq \pi_i$. Once again, this implies that $f = \pi_i$. In conclusion, f is a projection, and generalizing, F contains only projections and constants for $n \geq 2$. For $n = 1$, the only possible function other than the projection and the constants is negation, which cannot be in F since we would then have non-projections for $n \geq 2$. Then F contains only projections and constants, and is therefore a subset of **M** (notice we did not use WFA here).

The two remaining maximal proper subclones are the clone $\mathbf{P}_1$ of truth-preserving functions, functions such that $f(\vec{1}) = 1$, and the clone $\mathbf{P}_0$ of falsity-preserving functions, functions such that $f(\vec{0}) = 0$. If non-atomic STA and WFA are verified, A should contain some propositions that are true when all atoms are false, as well as propositions that are false when all atoms are true. It is therefore not possible for F to be a subset of $\mathbf{P}_0$ and $\mathbf{P}_1$. In conclusion, F is necessarily a subset of **M** if the A_n's are bilaterally weakly stable.

If STA and WFA proper are verified for all A_n, we know that F is a subset of **M**. There must be propositions that are true when all atoms are false, or false when all atoms are true: the only upward-monotonic functions staisfying this are the constants, which must then be in F. For $n = 2$, A must contain a proposition that is true when both atoms are true and stronger than both of them, so $\wedge$ must be in F, and similarly it must contain a proposition that is false when both atoms are false and weaker than both of them, so $\vee$ must be in F. Thus F contains a full basis of **M**, which means that $F = \mathbf{M}$.

A.3 Equivalence between the probabilistic notion of stability and STA

Communicative stability is defined in the context of a decision-theoretic model of pragmatics, in the spirit of the Rational Speech Act model (RSA, Goodman and Stuhlmüller 2013) or the Iterated Best Response model (IBR, Franke 2011). While both models we mentioned describe iterative procedures, we are only going to look at the first step.

We assume that there is a set of possible worlds W, a set of messages A which are propositions over W, and a prior probability distribution P_0 defined on W. A subset A_0 of the messages are called atoms. A listener is modelled by a probability distribution $L(\cdot|u)$ on W for each message u, interpreted as the posterior belief of the listener after having heard u. The literal listener L_0 performs Bayesian updates based on the fact that the message is true, using P_0 as a prior, as seen in (54).

(54) $$L_0(w|u;P_0) = P_0(\{w\} \mid [\![u]\!]) = \begin{cases} \frac{P_0(\{w\})}{P_0([\![u]\!])} & \text{if } w \in [\![u]\!] \\ 0 & \text{if } w \notin [\![u]\!] \end{cases}$$

The speaker is modelled by a function S mapping worlds to messages, interpreted as the message that speakers use in each situation. Relative to a certain model of the listener L, the utility of a message u for the speaker is written as $U(u|w;L,P_0)$ and is defined in (55). It quantifies the informativity of the message, which is also the posterior log-surprisal of the listener after hearing u. Models of the speaker are derived by optimizing a certain utility function: S and U are related through the equation in (56) for a certain L.

(55) $$U(u|w;L,P_0) = \log L(w|u;P_0)$$

(56) $$S(w;L,P_0) = \arg\max_u U(u|w;L,P_0)$$

Note that $U(u|w;L_0,P_0)$ is $-\infty$ when $w \notin [\![u]\!]$: when talking to the literal listener, false messages yield maximally low utility. This reflects Quality. The maximization in (56) reflects Quantity: speakers seek to make listeners maximally informed.[50]

Our main model of the speaker is the pragmatic speaker S_1, who is optimizing their speech with respect to the literal listener L_0; see (57). One can verify that as long as there are any true messages at a world, the pragmatic speaker always selects the most informative true message, that is, the message that had the smallest prior probability of being true among all the true messages.

(57) $$\begin{aligned} S_1(w;P_0) &= \arg\max_u U(u|w;L_0,P_0) \\ &= \arg\max_{u\in\mathbb{T}(w)} \log \frac{P_0(w)}{P_0([\![u]\!])} \\ &= \arg\min_{u\in\mathbb{T}(w)} P_0([\![u]\!]) \end{aligned}$$
where: $\mathbb{T}(w) = \{u : w \in [\![u]\!]\}$

Within this context, Bar-Lev and Katzir (2022a) define stability as in (58a), to which we add the definition of strict stability in (58b). Essentially, BL&K stability means that the production of non-atoms does not depend on the prior, and strict BL&K stability means that the production of no messages depends on the prior. If there is at least one true message at

[50]There is usually a cost term in U implementing Manner, but for our purposes it is better to consider that where Manner plays a role is in determining A.

every world, BL&K stability is equivalent to our weak stability, or to non-atomic STA, and strict BL&K stability is equivalent to our strong stability, or to STA.

(58) a. **BL&K stability:** A is said to be stable in the BL&K sense if at every world w, if for a certain setting of P_0 we have $S_1(w; P_0) = u^*$ for a certain $u^* \in A - A_0$, then any setting of P_0 is such that $S_1(w; P_0) = u^*$.

b. **Strict BL&K stability:** A is said to be strictly stable in the BL&K sense if at every world w, if for a certain setting of P_0 we have $S_1(w; P_0) = u^*$ for a certain $u^* \in A$, then any setting of P_0 is such that $S_1(w; P_0) = u^*$.

There are various corner cases in the above definitions that need to be addressed before we can embark with the proof. We assume that W is finite or countably infinite, so that we can write $P_0(w)$ (this is not crucial). A is assumed to be finite.[51] We assume that no worlds are ruled out by the prior, which makes the logarithms well-defined, and avoids potential problems with contextual equivalences.[52] We also assume that no two messages can have the same finite utility, which is sufficient to make our use of arg max well-defined. One reason this is reasonable to assume is that if P_0 is modelled as a continous random variable, then with adequate assumptions on how P_0 varies the probability of two particular messages being tied is 0. Finally, all choices of P_0 compatible with the above assumptions are possible, and they form a dense subset of the domain of valid probability measures.

We first prove that all minimal elements of $\mathbb{T}_A(w)$ can be optimal at w. Consider a world $w \in W$, and let p^* be a minimal element of $\mathbb{T}_A(w)$ (which exist and are finite in number because A is finite). Let P_0 be any valid prior. For $\alpha > 0$, let P_0^α be the distribution such that $P_0^\alpha(v) = \alpha P_0(v)$ if $v \in p^*$ and $P_0^\alpha(v) = \frac{1-\alpha P_0(p^*)}{1-P_0(p^*)} P_0(v)$ if $v \notin p^*$. One can straightforwardly verify that for any subset p of W, $P_0^\alpha(p)$ has limit $\frac{P_0(p-p^*)}{1-P_0(p^*)}$ as α tends to 0. In particular, if p is in $\mathbb{T}_A(w)$, because p^* is minimal, $p-p^*$ is non-empty, and it follows that $P_0^\alpha(p^*) < P_0^\alpha(p)$ for α small enough. Because there is a finite number of minimal elements, we can in fact find α such that $P_0^\alpha(p^*) < P_0^\alpha(p)$ for any p minimal in $\mathbb{T}_A(w)$ other than p^* itself. Take such an α. Because valid priors are dense, and because the number of inequalities is finite, we can find a valid prior P_0^* close enough to P_0^α that $P_0^*(p^*) < P_0^*(p)$ for any minimal p other than p^* itself. Now, because every proposition in $\mathbb{T}_A(w)$ is weaker that some minimal element of $\mathbb{T}_A(w)$ (again because A is finite), we in fact have $P_0^*(p^*) < P_0^*(p)$ for all $p \in \mathbb{T}_A(w)$ other than p^* itself. It follows that $S_1(w; P_0^*) = p^*$. Generalizing, all minimal elements of $\mathbb{T}_A(w)$ can be optimal at w for some prior.

The fact that all optimal elements are minimal is immediate. Then, the elements that can be optimal at w are exactly the minimal elements of $\mathbb{T}_A(w)$. The characterizations of stability and strict stability follow straightforwardly.

---

[51] I believe a weaker condition would suffice for the proof: the set of alternatives true at a given world always has a finite number of minimal elements. This condition would let us apply this result to infinite sets of alternatives based on discrete numeral scales. However, some proposals for infinite alternative sets based on continuous scales, such as that of Fox and Hackl (2006), are more fundamentally incompatible with utility-based models. In Fox and Hackl's analysis, worlds are essentially real numbers and propositions are open-ended intervals $]x; +\infty[$. The set of propositions true at x contains all $]y; +\infty[$ for $y < x$ and has no strongest elements, and there will generally not be a particular proposition whose utility is maximal at a given world. The analysis depends on this property and it probably cannot be recreated in a model based on numerical optimization.

[52] The problem is that if we know that Ann came, "Bill came" and "Both Ann and Bill came" perform the same update and we cannot distinguish them easily. Formal-logical models allow themselves to see the meaning of sentences beyond the current context and therefore have more possibilities in such cases.

Conflicts of interest

The author has no relevant conflicts of interests to report.

References

Abels, Klaus and Luisa Martí (2010). "A unified approach to split scope". In: *Natural language semantics* 18, pp. 435–470. DOI: 10.1007/s11050-010-9060-8.

Bar-Lev, Moshe E. and Danny Fox (2020). "Free choice, simplification, and innocent inclusion". In: *Natural Language Semantics* 28.3, pp. 175–223.

Bar-Lev, Moshe E. and Roni Katzir (2022a). "Communicative Stability and the Typology of Logical Operators". In: *Linguistic Inquiry*. DOI: 10.1162/ling_a_00497.

Bar-Lev, Moshe E. and Roni Katzir (2022b). "Positivity, (anti-)exhaustivity and stability". In: *Proceedings of the 23rd Amsterdam colloquium*. Ed. by Marco Degano, Thomas Roberts, Giorgio Sbardolini, and Marieke Schouwstra, pp. 23–30.

Barwise, Jon and Robin Cooper (1981). "Generalized quantifiers and natural language". In: *Linguistics and Philosophy* 4.2, pp. 159–219. DOI: 10.1007/BF00350139.

Bassi, Itai, Guillermo Del Pinal, and Uli Sauerland (2021). "Presuppositional exhaustification". In: *Semantics and Pragmatics* 14. DOI: 10.3765/sp.14.11.

Breheny, Richard, Nathan Klinedinst, Jacopo Romoli, and Yasutada Sudo (2018). "The symmetry problem: Current theories and prospects". In: *Natural Language Semantics* 26, pp. 85–110. DOI: 10.1007/s11050-017-9141-z.

Buccola, Brian, Manuel Križ, and Emmanuel Chemla (2022). "Conceptual alternatives". In: *Linguistics and Philosophy* 45, pp. 265–291. DOI: 10.1007/s10988-021-09327-w.

Chemla, Emmanuel, Brian Buccola, and Isabelle Dautriche (2019). "Connecting content and logical words". In: *Journal of Semantics* 36.3, pp. 531–547. DOI: 10.1093/jos/ffz001.

Chierchia, Gennaro (2004). "Scalar implicatures, polarity phenomena, and the syntax/pragmatics interface". In: *Structures and Beyond*. Ed. by Adriana Belletti. Oxford University Press, pp. 39–103.

Crnič, Luka (2011). "How to get "even" with desires". In: *Semantics and Linguistic Theory*, pp. 732–751. DOI: 10.3765/salt.v21i0.2600.

Crnič, Luka (2014). "Non-monotonicity in NPI licensing". In: *Natural Language Semantics* 22, pp. 169–217. DOI: 10.1007/s11050-014-9104-6.

Dayal, Veneeta (1996). *Locality in WH-quantification. Questions and relative clauses in Hindi*. Springer.

Elliott, Patrick D., Andreea C. Nicolae, and Uli Sauerland (2022). "Who and What Do *Who* and *What* Range Over Cross-Linguistically?" In: *Journal of Semantics* 39.3, pp. 551–579. DOI: 10.1093/jos/ffac008.

Enguehard, Émile and Emmanuel Chemla (2021). "Connectedness as a constraint on exhaustification". In: *Linguistics and Philosophy* 44.1, pp. 19–112. DOI: 10.1007/s10988-019-09286-3.

Enguehard, Émile and Benjamin Spector (2021). "Explaining gaps in the logical lexicon of natural languages. A decision-theoretic perspective on the square of Aristotle". In: *Semantics and Pragmatics*. DOI: 10.3765/sp.14.5.

Fox, Danny (2002). "Antecedent-Contained Deletion and the Copy Theory of Movement". In: *Linguistic Inquiry* 33.1, pp. 63–96. DOI: https://doi.org/10.1162/002438902317382189.

Fox, Danny (2007). “Free Choice Disjunction and the Theory of Scalar Implicatures”. In: *Presupposition and Implicature in Compositional Semantics*. Ed. by Uli Sauerland and Penka Stateva. Palgrave Macmillan, pp. 71–120.

Fox, Danny (2020). “Partition by Exhaustification: towards a solution to Gentile and Schwarz’s puzzle”. Ms. MIT (accessed April 2023). URL: https://semanticsarchive.net/Archive/TljZGNjZ/.

Fox, Danny and Martin Hackl (2006). “The universal density of measurement”. In: *Linguistics and Philosophy* 29.5, pp. 537–586. DOI: 10.1007/s10988-006-9004-4.

Fox, Danny and Roni Katzir (2011). “On the characterization of alternatives”. In: *Natural Language Semantics* 19.1, pp. 87–107. DOI: 10.1007/s11050-010-9065-3.

Fox, Danny and Roni Katzir (2020). “Notes on iterated rationality models of scalar implicatures”. Ms. MIT and Tel-Aviv University. URL: https://ling.auf.net/lingbuzz/005519.

Franke, Michael (2011). “Quantity implicatures, exhaustive interpretation, and rational conversation”. In: *Semantics and Pragmatics* 4.1, pp. 1–82. DOI: 10.3765/sp.4.1.

Gajić, Jovana (2019). “Negative coordination”. PhD thesis. Georg August University of Göttingen.

Giannakidou, Anastasia (1998). *Polarity sensitivity as (non)veridical dependency*. John Benjamins.

Goodman, Noah D. and Andreas Stuhlmüller (2013). “Knowledge and implicature. Modeling language understanding as social cognition”. In: *Topics in cognitive science* 5.1, pp. 173–184. DOI: 10.1111/tops.12007.

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

Heim, Irene and Angelika Kratzer (1998). *Semantics in Generative Grammar*. Blackwell.

Horn, Laurence R. (1973). “On the semantic properties of logical operators in English”. PhD thesis. University of California in Los Angeles.

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

Horn, Laurence R. (2012). “Histoire d’**O*: Lexical Pragmatics and the Geometry of Opposition”. In: *The Square of Opposition. A General Framework for Cognition*. Ed. by Jean-Yves Béziau and Gillman Payette. Peter Lang, pp. 393–426.

Incurvati, Luca and Giorgio Sbardolini (2022). “Update rules and semantic universals”. In: *Linguistics and Philosophy*, pp. 1–31.

Jacobs, Joachim (1980). “Lexical decomposition in Montague-grammar”. In: *Theoretical Linguistics*. DOI: 10.1515/thli.1980.7.1-3.121.

Just, Marcel Adam and Patricia Ann Carpenter (1971). “Comprehension of negation with quantification”. In: *Journal of Verbal Learning and Verbal Behavior* 10.3, pp. 244–253.

Katzir, Roni (2007). “Structurally-defined alternatives”. In: *Linguistics and Philosophy* 30, pp. 669–690. DOI: 10.1007/s10988-008-9029-y.

Katzir, Roni and Raj Singh (2013). “Constraints on the lexicalization of logical operators”. In: *Linguistics and Philosophy* 36.1, pp. 1–29. DOI: 10.1007/s10988-013-9130-8.

Keenan, Edward L. and Jonathan Stavi (1986). “A Semantic Characterization of Natural Language Determiners”. In: *Linguistics and Philosophy* 9.3, pp. 253–326. DOI: https://doi.org/10.1007/BF00630273.

Keenan, Edward L. and Dag Westerståhl (1997). “Generalized Quantifiers in Linguistics and Logic”. In: *Handbook of Logic and Language*. Ed. by Johan van Benthem and Alice ter Meulen, pp. 837–893. DOI: https://doi.org/10.1016/B978-044481714-3/50020-5.

Krifka, Manfred (1993). “Focus and Presupposition in Dynamic Interpretation”. In: *Journal of Semantics* 10.4, pp. 269–300. DOI: 10.1093/jos/10.4.269.
Ladusaw, William A. (1980). “On the Notion ‘Affective’ in the Analysis of Negative-Polarity Items”. In: *Journal of Linguistic Research* 1.2, pp. 1–16.
Meyer, Marie-Christine (2013). “Ignorance and grammar”. PhD thesis. Massachusetts Institute of Technology.
Penka, Doris (2010). *Negative indefinites*. Oxford University Press. DOI: 10.1093/acprof:oso/9780199567263.001.0001.
Percus, Orin (2006). “Antipresuppositions”. In: *Theoretical and empirical studies of reference and anaphora: Toward the establishment of generative grammar as an empirical science* 52, p. 73.
Peters, Stanley and Dag Westerståhl (2006). *Quantifiers in Language and Logic*. Clarendon Press.
Post, Emil L. (1941). *The Two-Valued Iterative Systems of Mathematical Logic*. Vol. 5. Annals of Mathematics studies. Princeton University Press.
Sassoon, Galit Weidman (2010). “The degree functions of negative adjectives”. In: *Natural Language sSemantics* 18.2, pp. 141–181. DOI: 10.1007/s11050-009-9052-8.
Sauerland, Uli (2004). “Scalar implicatures in complex sentences”. In: *Linguistics and philosophy* 27.3, pp. 367–391.
Sauerland, Uli (2008). “Implicated presuppositions”. In: *The discourse potential of underspecified structures*. Ed. by Anita Steube, pp. 581–600.
Sauerland, Uli (2012). “The computation of scalar implicatures: Pragmatic, lexical, or grammatical?” In: *Language and Linguistics Compass* 6, pp. 36–49. DOI: 10.1002/lnc3.321.
Schulz, Katrin and Robert van Rooij (2006). “Pragmatic meaning and non-monotonic reasoning: The case of exhaustive interpretation”. In: *Linguistics and Philosophy* 29.2, p. 205.
Schwarz, Bernhard (2017). “Negative polarity items: a case for questions as licensers”. In: *Semantics and Linguistic Theory*, pp. 230–247. DOI: 10.3765/salt.v27i0.4113.
Solt, Stephanie and Brandon Waldon (2019). “Numerals under negation: Empirical findings”. In: *Glossa: a journal of general linguistics* 4.1. DOI: 10.5334/gjgl.736.
Spector, Benjamin (2006). “Aspects de la pragmatique des opérateurs logiques”. PhD thesis. Université Paris-7.
Spector, Benjamin (2007). “Scalar Implicatures: Exhaustivity and Gricean Reasoning”. In: *Questions in Dynamic Semantics*. Ed. by Maria Aloni, Alastair Butler, and Paul Dekker. Brill, pp. 225–249. DOI: 10.1163/9780080470993_011.
Spector, Benjamin (2008). “An unnoticed reading for *wh*-questions: elided answers and weak islands”. In: *Linguistic Inquiry* 39.4, pp. 677–686.
Spector, Benjamin (2013). “Bare numerals and scalar implicatures”. In: *Language and Linguistics Compass* 7.5, pp. 273–294. DOI: 10.1111/lnc3.12018.
Spector, Benjamin (2016). “Comparing exhaustivity operators”. In: *Semantics and Pragmatics* 9, pp. 11–1.
Spector, Benjamin and Yasutada Sudo (2017). “Presupposed ignorance and exhaustification: how scalar implicatures and presuppositions interact”. In: *Linguistics and Philosophy* 40.5, pp. 473–517. DOI: 10.1007/s10988-017-9208-9.
Steinert-Threlkeld, Shane and Jakub Szymanik (2019). “Learnability and semantic universals”. In: *Semantics and Pragmatics* 12. DOI: 10.3765/sp.12.4.
Szymanik, Jakub (2016). *Quantifiers and cognition: Logical and computational perspectives*. Springer.

Uegaki, Wataru (2022). “The Informativeness/Complexity Trade-Off in the Domain of Boolean Connectives”. In: *Linguistic Inquiry*. DOI: 10.1162/ling_a_00461.

Von Fintel, Kai (1999). “NPI licensing, Strawson entailment, and context dependency”. In: *Journal of semantics* 16.2, pp. 97–148. DOI: 10.1093/jos/16.2.97.

Wurmbrand, Susi (2008). “*Nor*: Neither Disjunction nor Paradox”. In: *Linguistic Inquiry* 39.3, pp. 511–522.

Xiang, Yimei (2021). “Higher-order readings of *wh*-questions”. In: *Natural Language Semantics* 29, pp. 1–45. DOI: 10.1007/s11050-020-09166-8.

Zeijlstra, Hedde (2004). “Sentential negation and negative concord”. PhD thesis. University of Amsterdam.