Some formal results on the upward-monotonic bias, communicative stability, the strongest answer condition, and exhaustification

Émile Enguehard
ILLC, University of Amsterdam

Outline

The puzzle that this paper is attempting to address is the following: why does natural language seem to give privileged status to the monotonic, and particularly upward monotonic Boolean-valued functions? We will begin by defining monotonic functions, as well as review a variety of puzzling cases in which semantic theory appears to need to appeal to a preference for such functions over other logical functions; this is the topic of Section 1. In Section 2, we will introduce the notion of communicative stability, a property defined in the context of decision-theoretic models of pragmatics and put forward by Bar-Lev and Katzir (2022a) to account for certain of the aforementioned puzzles. Stability essentially states that the structure of the set of messages available to pragmatic speakers — speakers who are modelled as optimizing a certain utility function — is such that their predicted behavior does not depend on prior probabilities. We will prove that stability can be characterized in non-probabilistic terms and is essentially equivalent to a strongest answer condition (as per Dayal 1996) applied to alternative sets. On this basis, we will offer an alternative justification for the desirability of stability in the linguistic system: an alternative set is stable if and only if applying an exhaustification operator to the set of messages produces the best possible partition of the logical space. Finally, in section 3, we will prove several results to the effect that restricting the language to upward-monotonic functions is necessary or optimal to achieve stability, as long as we take the presence of atomic propositions for granted, with a variety of ways of filling out the details. Inasmuch as stability is desirable, this can be seen as an explanation for the predominance of upward-monotonic operators in natural language.

Acknowledgements

I thank Benjamin Spector, Moshe E. Bar-Lev and Roni Katzir for discussion and for encouraging me to write it all. All errors are the author's. 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*.

1 Background: the ubiquity of monotonicity in natural language

1.1 Definitions and terminology

A function f between two partially ordered sets is said to be *upward-monotonic* whenever it satisfies (1a), and *downward-monotonic* whenever it satisfies (1b).[1]

(1) a. $\forall x, y.\, x \leq y \rightarrow f(x) \leq f(y)$
b. $\forall x, y.\, x \leq y \rightarrow f(y) \leq f(x)$

In the specific case where f ranges over truth values, these conditions can be rewritten as follows:

(2) a. $\forall x.\, f(x) \rightarrow \forall y \geq x.\, f(y)$
b. $\forall x.\, f(x) \rightarrow \forall y \leq x.\, f(y)$

Or, identifying f to the set for which it is the characteristic function:

(3) a. $\forall x.\, x \in f \rightarrow \forall y \geq x.\, y \in f$
b. $\forall x.\, x \in f \rightarrow \forall y \leq x.\, y \in f$

Thus, traveling "up" in a set whose characteristic function is upward-monotonic, we will never evit the set. For this reason, a set whose characteristic function is upward-monotonic is said to be *upward-closed*, while a set whose characteristic function is downward-monotonic is *downward-closed*.

We are going in a few places to use the related notion of convexity, discussed under the name "connectedness" by Chemla et al. (2019) and Enguehard and Spector (2021).[2] A logical function is convex if its truth set does not have "holes" in it: falsity worlds that are in-between truth worlds with respect to the order. A formal definition is given in (4). One may easily verify that all monotonic functions are convex.

(4) **Convexity:** a logical function F is convex if it is true at any input *in-between* two other inputs where it is true. Formally, for any three inputs $\vec{p}_1$, $\vec{p}_2$ and $\vec{p}_3$ such that $\vec{p}_1 \leq \vec{p}_2 \leq \vec{p}_3$, if $F(\vec{p}_1)$ and $F(\vec{p}_3)$ are true then $F(\vec{p}_2)$ is true.

A concrete example of these notions can be given by considering the specific case of generalized quantifiers, that is, functions of type $(et)t$. The domain of these functions is objects of type et, the type of sets of entities, or predicates. There is a natural partial order on

[1] The dominant convention in certain formal fields of mathematics is to use *monotone* or *monotonic* to refer to what we call here upward-monotonic functions, while downward-monotonic functions are said to be *antitone* or *antitonic*. Another, more descriptive possibility is to use *order-preserving* and *order-reversing*. Either of these choices leaves us without a word to refer to what we call monotonic functions, even though this notion is often useful when talking about natural language, which is among the reasons why we adopt the upward/downward terminology instead. This terminology also has the advantage, in the case of Boolean-valued functions, of emphasizing the connection with upward- and downward-closed sets, and of being visually intuitive when one looks at a graphical representation of the order. This why we prefer it to *monotone increasing* and *monotone decreasing*, which are generally used when talking about real-valued functions.

[2] 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 topology and related branches 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 it, while a connected set is a set where there is always a way to go from one element from another.

predicates given by the relation of set inclusion. The quantifier **everybody** given in (5) (the denotation of the English word *everybody*, or close enough) can be verified to be upward-monotonic: if $P \subseteq Q$, then **everybody** P entails **everybody** Q. In contrast, the quantifier **nobody** given in (6) (the denotation of *nobody*) is downward-monotonic.

(5) $\mathbf{everybody} = \lambda P.\, \forall x.\, P(x)$

(6) $\mathbf{nobody} = \lambda P.\, \neg\exists x.\, P(x)$

The notion of *upward-entailing* (UE) and *downward-entailing* (DE) contexts found in the semantics literature and usually defined in terms of entailment relaitons between sentences is closely linked to that of upward- and downward-monotonicity. If we consider a sentence $\varphi[\_]$ with a "hole" in a certain syntactic position, such that the hole can be filled with elements of a certain type τ that ends in t (and is, therefore, naturally ordered through set inclusion), then we can define a Boolean-valued function that maps an element x of type τ to $[\![\varphi[\theta]]\!]$, where θ is such that $[\![\theta]\!] = x$. Then, the context of the position is said to be upward-entailing whenever the associated function is upward-monotonic, and downward-entailing when the function is downward-monotonic. For instance, the position of the VP in a sentence whose subject is "everybody", or in other words the hole in "Everybody [_]", maps to the function **everybody**, and is therefore an upward-entailing context, while the VP-shaped hole in "Nobody [_]", which maps to the function **nobody**, is a downward-entailing context. Thus we have three closely related pairs of words:

Kind of objects	Boolean-valued functions	Sets	Syntactic positions
upward- / downward-	*monotonic*	*closed*	*entailing*

Phrases like *somebody*, *at least two people*, or *many people* are analyzed as denoting upward monotonic quantifiers, while phrases like *few people* or *at most two people* are analyzed as denoting downward-monotonic quantifiers; meanwhile the denotation of *exactly two people* is not monotonic in either direction, but it is convex. An example of a non-convex quantifier is *many people or nobody* (inasmuch as English allows this phrase). One can get a visual intuition for all these facts by looking at a graphical representation of the order on predicates.

As an interesting special case, a proper noun like *Ann* in subject position creates an upward-entailing context for the VP position. The function corresponding to the gappy sentence "Ann [_]" is indeed the lifted denotation of *Ann*, which we write as $^{\Uparrow}\mathbf{a}$ (where $\mathbf{a}$ is understood to be the non-lifted, type e denotation of *Ann*) and give in (7). This quantifier is upward-monotonic.

(7) $^{\Uparrow}\mathbf{a} = \lambda P.\, P(\mathbf{a})$

Thus there is an intuition that upward-entailing contexts are the default kind of contexts, since they are observed in the simplest sentences. The results we will present in Section 3 can be seen as showing that, given that we start with upward-monotonicity in the most basic sentences, there are benefits to the logical lexicon being such that more complex sentences also will involve upward-monotonicity.

Both upward- and downward-monotonicity are closed under conjunction and disjunction, while negation reverses the direction of monotonicity.[3] This is a specific instance of the fact that upward- and downward-monotonicity are closed under composition with

[3]Where we define conjunction, disjunction and negation over all types that end in t in the natural way.

upward-monotonic functions, while composition with downward-monotonic functions reverses the direction; indeed, conjunction and disjunction (seen as functions over pairs ordered in pointwise fashion) are themselves upward-monotonic operations, while negation is downward-monotonic. From the above, we can deduce that the negation of a monotonic function is convex (because it is monotonic).[4] Convexity is not closed under negation in general, nor under disjunction, but it is closed under conjunction.

In the rest of this section, we will review the different ways in which monotonicity shapes the linguistic system. What we will argue is that some kind of preference for upward-monotonic expressions, which we will call the upward-monotonic bias, can be diagnosed in relation to a variety of syntactic, semantic and pragmatic phenomena.

1.2 Monotonicity in syntax: negative polarity items

Probably the most famous appearance of the notion of monotonicity in formal semantics is its usage to characterize the distribution of negative polarity items. These are items such as English *anybody* (and other items containing *any*) which are not licensed in simple positive sentences; environments where they are licensed include the scope of negation, the restrictor of universal quantifiers, the scope of downward-monotonic quantifiers, the scope of exclusives, as well as inside temporal or conditional adverbial clauses. The common trait of the syntactic contexts we just listed is that they are downward-entailing, which leads us to the so-called Fauconnier-Ladusaw hypothesis (Ladusaw 1980), according to which downward-entailingness is in fact the explanatory criterion for NPI licensing.[5]

Of course, we want to in turn explain this association. It has been proposed that NPIs have a domain-widening effect that allows speakers to trigger certain scalar implicatures, in a way that does not work out in UE contexts (Kadmon and Landman 1993, a.o.), that they are used for their alternatives, again to trigger certain scalar implicatures in a way that does not work out in UE contexts (Krifka 1995; Chierchia 2013, a.o.), or that they associate with presuppositional operators so as to trigger certain inferences about speaker attitudes, in a way that does not work out in UE contexts (Chierchia 2013; Crnič 2014a, a.o.). The common trait of these proposals is that NPIs' "function" is to trigger pragmatic inferences, and that what makes most NPIs behave like an NPI is the fact that they are "minimizers" at the bottom of a scale, that is, the weakest expression among a set of expressions. For instance, *any student* is seen as weaker than any other expression quantifiying over students.[6]

What none of these proposals explain is why NPIs exist, or in other words, why natural languages feel the need to have elements with such properties and mark them in a special way. One can imagine marked "maximizer" elements which would trigger pragmatic inferences of a symmetric kind to those triggered by NPIs, in a way that only makes sense in non-DE contexts, by virtue of the fact that they are on top of a logical scale. While so-called positive polarity items have indeed been identified, if we look at the list of such elements given by Spector (2014), we see that they are mostly "weak" items (disjunctions, existentials, etc., including *some* in English), or words expressing some sort of approxima-

[4] A sufficient condition for the reverse implication to hold (that if a function and its negation are convex, it is monotonic) is that the domain is a bounded lattice.

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

[6] Interestingly, this statement presupposes that *any* expressions are not compared with any downward-monotonic quantifiers. The precise class of quantifiers that it allows is those that Xiang (2021) calls "positive".

tion (*almost*, *about*...) and therefore not what we are looking for here. Why do we not have a large inventory of universal-y PPIs like we have a large inventory of existential-y NPIs?

We are not going to really explain it here, but we will simply relate it to the general intuition that the linguistic system takes UE-ness to be the basic case, and DE-ness to be marked: thus the words that speakers use in UE contexts are the default ones for a given meaning, and do not trigger as many inferences or have such a restricted distribution as the words used in DE contexts. Now, one might object that this is *ad hoc*, and that we could rationalize the opposite situation in a similar way, and of course the opinions of the linguistic system is not a very serious notion (and is it not downward-entailingness that the system prefers, since it lets us use more words in that case?). The important, undeniable thing, is that a sizable section of the logical lexicon is made of items whose distribution tracks monotonicity, and that both directions are not interchangeable.

1.3 Monotonicity in the lexicon: Horn's puzzle and related topics

We have already seen how the notion of monotonicity can be applied to the case of generalized quantifiers. In their foundational work on the topic, Barwise and Cooper (1981) propose certain universals relating to monotonicity, which we can restate in our own words here:

(8) a. That if a downward-monotonic quantifier is expressible in a simple way in a language, so is its (upward-monotonic) negation (**U5**);
b. That if a quantifier is expressible in a simple way in a language, then at the semantic level it is expressible as a conjunction of monotonic quantifiers (**U6**).

What "simple" means here is essentially that this generalization bears on quantifiers proper (such as *everybody*), quantifying determiners, and modified numerals; coordinated structures and expressions involving overt negation do not count. U6 rules out that a simple expression in this sense could be synonymous, for instance, with *all or none*. It does not completely rule out non-monotonic meanings; non-monotonic convex expressions, which includes in particular synonyms of *some but not all*, are allowed, even though it seems that these can also be ruled out if we focus on one-word expressions.[7] U6 gives a special role to monotonicity, but does not differentiate the directions. U5, however, gives a privileged role to upward-monotonicity: it says, essentially, that there are more upward-monotonic quantifers than downward-monotonic ones. This fact, of course, begs for an explanation.

As examples of upward-monotonic quantifying expressions that do not have a lexicalized negation, we can mention *most* and *all*. The specific fact that no single word **nall* meaning *not all* exists is in fact the subject of a sizable body of literature, going back to an observation due to Horn (1973, chap. 4). Horn's observation is framed in relation to a traditional four-way taxonomy of quantified statements known as the Square of Opposition or Square of Aristotle. Each category is labelled with one of the latters A, E, I and O, as follows:

(9) a. Universal affirmative (A): *All men are mortal.*
b. Universal negative (E): *No men are mortal.*
c. Particular affirmative (I): *Some men are mortal.*
d. Particular negative (O): *Not all men are mortal.*

[7] For the relevance of convexity to the logical lexicon of natural languages, see Chemla et al. 2019 and Enguehard and Chemla 2019.

Note that the distinction between the words "affirmative" and "negative" here, is about the direction of monotonicity of the initial determiner. These four kinds of statements are logically related, in such a way that there is systematic symmetry between the affirmative A and I and the negative E and O: A asymmetrically entails I just as E asymmetrically entails O (in both cases, assuming existential import), O is the negation of A just as I is the negation of E, and A and I are dual just as E and O are dual. Horn's observation is essentially that this symmetry is not respected by the lexicon of natural languages, and that the upward-monotonic A and I tend to be more integrated into the lexicon in various ways.

The reason Horn's observation is not just a special case of Barwise and Cooper's is that it extends to other families of logical functions comprising a "strong" upward-monotonic function (A), its weaker dual (I), and their respective negations (O and E). Table 1 lists a number of English examples; see Horn 1973, chap. 4 for an extensive discussion, as well as Horn 2012 and references therein. What we observe, in general, is a kind of hierarchy: $A, I > E > O$. While A and I are always lexicalized, E is often missing, and O almost always. Additionally, when E and O are lexicalized, they are almost always morphologically more complex than A and I — in fact, they are often derived from A and I with the addition of a morpheme expressing negation — and O is often more complex or more clearly derived than E.[8]

What is of interest to us here is that the two elements on top of this hierarchy, A and I, are precisely the two upward-monotonic corners of the square.

A	I	E	O
all	*some*	*no*	*(*nall)*
and	*or*	*(nor)*[9]	*(*nand)*
both... and...	*either... or...*	*neither... nor...*	—
always	*sometimes/ever*	*never*	—
necessary	*possible*	*impossible*	*unnecessary*[10]
must	*can/may*	*(cannot)*	*(needn't)*[11]

Table 1: Examples of logical squares

[8] In this discussion, we take for granted that elements like *nobody* and counterparts in other languages, which create downward-entailing contexts, are in fact lexical downward-monotonic operators, so that natural language does have downward-monotonic operators. There are at least two reasons to challenge this assumption. First, the possibility of split-scope readings for such operators suggests that they are morphosyntactically complex, even in languages where their surface form does not make it apparent (see Abels and Martí 2010 and references therein). Second, in many languages, some of these operators exhibit negative concord: when several such operators occur together, there is "only one negation" in a certain sense, suggesting they incorporate a separable clause-level negation or even that they systematically co-occur with such a negation without it being part of their meaning, with the remaining component of the meaning being upward-monotonic (see Zeijlstra 2004; Kuhn 2022 a.o.). The analyses that have been developed to account for either of these facts take us closer to a picture where clause-level negation is the only downward-monotonic operator in natural language, with all the rest of the logical lexicon being upward-monotonic; if this is correct, it of course strengthens our argument.

Given the fact that natural languages can express negation, and the logical links between the four corners, it is sufficient to lexicalize just one corner to express all four, and any of them will do. There is then no obvious reason why we generally observe two or three operators, nor why A and I should be privileged. Horn remarks that, putting the possibility of negation aside, I and O statements can be used to express the other thanks to scalar implicatures, a point to which we will come back; this can be seen to explain why we rarely see a four-element lexicon — even though speakers do use O statements[12] — but not why I is preferred to O within the two weak corners, nor why 2-element lexicons are common.

One can describe Horn's observation and Barwise and Cooper's U5 as the fact that negation is marked, or that negative elements are derivative of positive ones, but inasmuch as negative elements are identified from their morphology, and that markedness and derivation are also diagnosed on the basis of morphology, there is something a bit circular to it. We can instead see it as a dispreferrence towards non-upward-monotonic operators; this is not an explanation either, but it links the observation to a more general pattern. Besides, as Katzir and Singh (2013) point out, this lets us also capture the fact that the dispreferrence extends to non-monotonic operators, and not just in the case of quantifiers: there is no binary connective meaning an exclusive disjunction or a bi-implication, etc.

Katzir and Singh (2013) propose that logical operators' meanings are represented in a certain language whose primitive constructs are such that upward-monotonic functions are generally simpler; a similar proposal is made by Uegaki (2022), building up on Katzir and Singh's idea. Such proposals let us displace the question again, but do not fundamentally solve it: instead of explaining why upward-monotonicity is simpler, we have to explain why the representation language is the way it is. Another proposal of this kind, from a fairly different conceptual perspective, is made by Incurvati and Sbardolini (2022) for the specific

[9]It is sometimes said that English *nor* expresses negated disjunction (logical NOR), but attested examples are exceptional (see Horn 2012 for some of them). Some other items like Dutch *noch* do seem equivalent to logical NOR in their unembedded binary use, see (ia), but this equivalences breaks down in ternary uses or embedded cases, see (ib) and (ic).

(i) a. Jan drinkt noch eet.
Jan neither drinks nor eats."
b. Jan drinkt noch eet noch slaapt.
"Jan neither drinks, nor eats, nor sleeps." (not: "Jan drinks or sleeps, and doesn't sleep.")
c. Niemand drinkt noch eet.
"Nobody drinks and nobody eats." (not: "Everyone drinks or eats.")

Instead, *nor* and its counterparts 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).

[10]Horn considers *unnecessary* to be "less lexicalized" than *impossible* because it makes use of a productive negative prefix. He also notes that unlike the previous three adjectives which can be epistemic, *unnecessary* only has a deontic reading.

[11]Horn makes the interesting point that A-corner modals often show neg-raising behavior, which means that their negation expresses E and not O; English *must* is an example. We can extend the observation to other cases of "gappiness", such as the fact that negating plural definites, which express A, expresses E rather than O. These seemingly unrelated phenomena conspire to make O harder to say. It seems plausible that the dispreferrence for O has a role to play in the analysis of neg-raising; Jeretič 2022 is one proposal that incorporates this idea.

Another remarkable fact is that in many languages, a non-neg-raising A-corner modal is specialized for O and chiefly used with negation; English *need* (as in *needn't*) and Dutch *hoeven* are two examples. This shows that O is a useful enough thing to say that dedicated expressions can arise; yet even in this case, the preference is for these expressions to incorporate a negation of some kind.

[12]See Enguehard and Spector 2021 for a discussion of the use conditions of O statements.

case of binary connectives; they relate the absence of **nand* to consideration of conversational dynamics. Again, this line of explanation depends on a specific set of representational primitives, which one needs in turn to explain. Enguehard and Spector (2021) escape this problem by deriving the dispreferrence for O from an independent (in the logical sense) generalization about the denotation of NPs; however, they account solely for the choice between the two lexicons A, I, E, and A, O, E, and it is not clear that a general preference for upward-monotonicity follows from their results. Thus the question remains largely unsolved. We will of course come back to the proposal of Bar-Lev and Katzir (2022a), which this paper builds up on.

1.4 Monotonicity in pragmatics: the symmetry problem in the derivation of implicatures

Another reflex of the monotonicity bias can be seen in the so-called "symmetry problem" in the theory of implicature derivation (see Breheny et al. 2018 for a longer discussion). We are going to illustrate this problem using the categories of quantified statements of the square of Aristotle, with the addition of a two-sided particular statement OI, as in (10).

(10) Two-sided particular (OI): *Some but not all men are mortal.*

The symmetry problem, essentially, is the fact that the existential I statement appears to have among its alternatives A but not OI or O. Indeed, we generally observe from the utterance of I an implicature that O is true, which can be explained straightforwardly in terms of Gricean reasoning relative to the alternative A, as implemented for instance by Sauerland (2004). Essentially, if the speaker had believed A to be true, since it is more informative than I, they should have used it instead. However, if we apply this reasoning to OI, then we predict that I should imply that A is true, contrary to fact. If both alternatives are considered at the same time, most theories predict no implicature, again contrary to fact. Hence A must be an alternative and OI must not be. Additionally, if we consider that alternatives need not be strictly stronger than the utterance, we also have to exclude O for the same reasons as OI.

The problem is that from a purely logical point of view, there is no difference between A and OI in terms of their relation to I: both are logically stronger than I, and they partition the set of worlds compatible with I. They are referred to as "symmetric" alternatives for this reason. The symmetry problem is the problem of distinguishing them in our theory of alternatives.

Fox and Katzir (2011) propose that alternatives are obtained by replacing elements of the uterrance by alternative elements of equal or lower syntactic complexity. In OI, the initial determiner involves a coordinated structure, and is clearly more complex than the one-word initial determiner in I and A, which explains why OI is not an alternative to I. O is excluded as well for similar reasons. Notice how this explanation is dependent on the fact that no simple way to express OI or O exists. Thus it is the upward-monotonic bias which breaks the symmetry: because O and OI 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.[13] The symmetry problem can therefore be seen as a reflex of the upward-monotonic bias.

[13] In the conclusion, I will submit that on the basis of the results presented in this paper, this might in fact be the *function* of the upward-monotonic bias.

The symmetry problem also occurs for indirect implicatures, that is, implicatures from negative statements. As Breheny et al. (2018) discuss, the case of indirect implicatures is potentially problematic for Fox and Katzir's theory. If we negate everything, we get a very similar situation to the one we have discussed above: uttering O implicates the truth of OI and I as well as the falsity of E, which suggests that E is an alternative, but not OI or I. We can argue as before that OI is more complex, but we have a problem if alternatives that are not stronger than the utterance are possible, as they are in some theories. Then, Fox and Katzir's theory fails to exclude I, as it is not more complex than E or O. A simple but stipulative solution, at least for this particular case, is to assume that logical expressions may only be replaced by expressions of the same monotonicity to form alternatives.[14] Whatever the exact solution, the case of indirect implicatures reinforces the conclusion that considerations of monotonicity shape alternative sets, beyond the question of the direct logical relation between the utterance and its alternatives.

1.5 Monotonicity in the semantics of questions

Beyond implicature derivation, the symmetry problem also affects the semantics of focus-sensitive items like *only* (Fox and Katzir 2011). Closely related issues also occur in the semantics of questions, as we will briefly illustrate here.[15]

A common paradigm in question semantics is the answer set theory, which sees questions as sets of propositions, corresponding intuitively to the possible answers. In systems inspired by Hamblin 1976, a question directly denotes the set of its possible answers, which is obtained by replacing the *wh*-word with elements of its domain.[16] A typical analysis of a basic *wh*-question are would be what is given in (11): the *wh*-word is replaced by all relevant type *e* elements to obtain the answer set.

(11) $[\![\text{Who came?}]\!] = \{\text{“Mary came”, “John came”, “Mary and John came”}\}$

The answer set in (11) does not include any “negative” propositions like *Mary did not come*. This follows straightforwardly from the method of composition, and seems unremarkable from the historical perspective of answer set semantics. I would nevertheless argue that it is not self-evident: one might expect that answerhood could be reduced to some notion of relevance, and in turn that relevance should be closed under logical operations like negation,[17] so that negative propositions should be in the set along with positive ones. This would be problematic for at least two reasons. First, in question-answer pairs an inference that the given (positive) answer is the most informative true answer is reliably observed; cf. (12). This inference can straightforwardly be analyzed as a scalar implicature, as long as the alternative set contains only other positive answers — otherwise, we run into the symmetry problem. If we want to unify the notions of answer set and alternative set, we then need to keep the negative answers outside of the question denotation. Second, there exist so-called “mention-some” questions (as opposed to “mention-all”) where incomplete

[14] If we accept that downward-monotonic operators are always underlyingly formed of an upward-monotonic operator and a negation, our stipulation would come down to the fact that negation cannot be deleted to form alternatives. Alternatively, it can be seen as a variation on Horn scales (Sauerland 2004 and references therein), with an explicit constraint.

[15] This section draws heavily from [author's PhD thesis].

[16] The alternative is systems inspired by Karttunen 1977 where a question denotes the sets of its true answers at a given world; the difference does not matter for our purposes.

[17] If relevance is defined in terms of Groenendijk and Stokhof's (1984) partition theory of questions, it is in fact closed under logical operations.

positive answers are acceptable, as in (13a). Yet not any relevant information makes for a good mention-some answer; in particular, negative answers such as that in (13b) are judged to be incomplete. This is impossible to account for if negative answers are in the denotation.

(12) Q: Who knows the password?
A: Mary does.
⇝ John does not know the password.

(13) Q: Who can tell me the password?
a. A: Mary can. ⇝̸ John does not know the password.
b. #A: Mary cannot.

All in all, it is a theoretical necessity for the characterization of good mention-some answers, and for explaining the strong interpretation of mention-all answers, that the answer set should be such that it does not contain any negative answers. While I acknowledge that this is only one possible framing, we could relate this fact to the other phenomena we discuss here by characterizing it as a restriction to upward-monotonic elements in building up the denotation of a question — recall that proper nouns, seen as quantifiers, are upward-monotonic. We then have yet another instance of the upward-monotonic bias.

More direct evidence for the existence of a restriction to upward-monotonicity in question denotations is found when we look at questions where the *wh*-word ranges over higher-type elements. Spector (2008) argues for the existence of such questions as well as the fact that their answer set is restricted to upward-monotonic quantifiers. 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 (14), (14a) appears to entail (14b) but to be compatible with the falsity of (14c). This is readily explained if the *wh*-phrase in (14) ranges over upward-monotonic quantifiers only.[18]

(14) 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*.

1.6 Monotonicity in the mind

To conclude our review of the upward-monotonicity bias, let us briefly mention that it has also been claimed to be observable outside of language proper, in psychological studies, though the evidence is limited. In experiments conducted by Geurts and van Der Slik (2005), participants were tasked with assessing the validity of inferences between sentences involving quantifying expressions (similar to syllogistic reasoning). Geurts and van Der Slik find that participants make fewer errors when the sentences are based on upward-monotonic quantifiers than when downward-monotonic quantifiers occur. Meanwhile, Chemla

[18] See also Fox 2020. More recently, Xiang (2021) has argued that the restriction is in fact to positive quantifiers, that is, quantifiers that entail an existential statement. This includes all non-trivial upward-monotonic quantifiers and excludes all non-trivial downward-monotonic ones, but it also tolerates some non-monotonic ones.

et al. (2019) find that participants in a rule-learning task learn the rule faster if the rule involves monotonic operators than if it involves non-monotonic ones. Note that, in Chemla et al.'s experiment, no linguistic stimuli or productions are involved. This suggests that monotonic functions are easier to process in some way, and that within those functions, upward-monotonic ones are easier than downward-monotonic ones. Then, the upward-monotonic bias in language would be linked to a cognitive bias in the usual sense of the word.

It is tempting to see, for instance, the lexicalization facts as a consequence of the cognitive bias: we do not lexicalize logical concepts that are hard to think about. However, there is a plausible causal pathway in the other direction: inasmuch as natural language is used by its speakers to help their thought, concepts that are harder to express should also be harder to think about. For this reason, the existence of a cognitive bias is not the end of the story. On top of that, as we have already mentioned, 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 I / *some* or O / *not all* will be interpreted as OI / *some but not all*, as we have described above. This seems hard to reconcile with the idea that non-monotonic or downward-monotonic operators are so challenging to process that lexicalization is impossible.

2 Communicative stability: statement and properties

The contribution of this paper is to build up on a proposal made by Bar-Lev and Katzir (2022a) to account for Horn's puzzle. Horn's puzzle, as discussed in Section 1.3, has to do with the absence of the O corner of the Square of Opposition in the logical lexicon, and more generally the predominance of upward-monotonic operators. Bar-Lev and Katzir propose that the lexicon is shaped so that the alternative sets entering pragmatic computation have a certain formal property they call *communicative stability*, or just *stability* for short. This property guarantees that the behavior of speakers, as predicted by certain probabilistic models, does not depend on the prior probability parameter of the models in question.

Bar-Lev and Katzir show through exhaustive enumeration that, in the case of binary connectives, only upward-monotonic connectives are safe to lexicalize if one wants to preserve stability together with a few other natural requirements. In the next section, we are going to present a few generalizations or variants of their results in an effort to extend the idea to arbitrary Boolean functions. Before we do this, we have to explain what stability is; this is the object of this section. We are also going to show that stability is equivalent to Dayal's (1996) strongest answer condition, applied to alternative sets, and relate it to the effect of exhaustification on the alternative set; we will argue that it provides additional motivation for seeing stability as desirable.

2.1 Decision-theoretic approaches to implicature generation

Communicative stability is defined in the context of a decision-theoretic model of pragmatics. Such models include the Rational Speech Act model (RSA, Goodman and Stuhlmüller 2013) or the Iterated Best Response model (IBR, Franke 2011), and have been used in particular to account for scalar implicatures. While both models we mentioned describe iterative procedures, we are only going to look at the first step. We are conceptually close to RSA in

that we assume the prior probability parameter truly corresponds to some kind of Common Ground epistemic state which varies with situations, but formally close to IBR in that our procedure is deterministic. Note that the model we describe here is roughly the same one as that of Enguehard and Spector (2021) as well as Bar-Lev and Katzir (2022a).

Because we are ultimately interested in proving formal results, we need to be precise about our assumptions. Nevertheless, to make the exposition easier to read, we will keep certain definitions incomplete and certain assumptions unstated in our general description of the model. At the end of this section, under the label "Technical notes", a series of addendums and clarifications can be found which (hopefully) makes the setting for our proofs well-specified.

Let us assume, then, that there is a set of possible worlds W, a set of messages A, each member u of which is associated to its semantics $[\![u]\!] \subseteq W$, and a prior probability distribution P_0 defined on W. We will call A the alternative set; it is important not to call it the lexicon since we are going to use that term for sets of words, and A is a set of sentences. 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. In particular, we can define the literal listener L_0, who performs Bayesian updates based on the fact that the message is true, using P_0 as a prior, as seen in (15). Conceptually, L_0 accepts what they hear to be true (as per the Maxim of Quality), but does not perform any extra pragmatic reasoning.

$$L_0(w|u;P_0) = P_0(\{w\} \mid [\![u]\!]) = \begin{cases} \frac{P_0(\{w\})}{P_0([\![u]\!])} & \text{if } w \in [\![u]\!] \\ 0 & \text{if } w \notin [\![u]\!] \end{cases} \tag{15}$$

The speaker is assumed to know exactly which world we are in. They are 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 (16). 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 (17) for a certain L.

$$U(u|w;L,P_0) = \log L(w|u;P_0) \tag{16}$$

$$S(w;L,P_0) = \arg\max_u U(u|w;L,P_0) \tag{17}$$

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 (17) reflects Quantity: speakers seek to make listeners maximally informed.[19]

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 (18). 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.

[19]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.

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

Conceptually, the fact that S_1 speaks to L_0 implements the idea that speakers obey the Gricean maxims and that P_0 is the Common Ground prior. Since listeners know that speakers obey the Gricean maxims and know what the Common Ground is, they can also calculate the function S_1; then, by inverting it, they can get information about the world we are in that is possibly more precise that the literal meaning of the message.[20] This is how our model can predict pragmatic reasoning.

Technical notes This part of the section is concerned with specifying the model more properly to make it useable in proofs. In particular, we need to address various corner cases of the definition of S_1. The assumptions that we make are below. In several cases, one of these assumptions subsumes another, but we still list both separately for conceptual reasons.

1. W **is finite or countably infinite.** This assumption is necessary to let us write $P_0(w)$, but not crucial, and we would only have to complicate everything a bit if we wanted to drop it.

2. A **is finite.** This lets us take maximal and minimal elements under entailment within arbitrary subsets of A. It also guarantees the existence of optimal messages in terms of utility.[21]

3. **No worlds are ruled out by the prior.** While we allow priors to vary, we assume that the context set (the support of the prior) remains fixed, and we assume that W is restricted to the context set. First, this makes our logarithms well-defined. Second, without this assumption, certain choices of prior that assign probability 0 to certain regions of logical space would make some messages become contextually equivalent, which is generally problematic in models inspired by Bayesianism.[22]

[20]The assumption that participants know P_0, or even that they know A and the semantics $[\![\cdot]\!]$, is not crucial. If we drop it, we can model how participants get information about (the speaker's beliefs about) P_0, A and $[\![\cdot]\!]$ from the speaker's choice of message.

[21]I believe a weaker condition would suffice for the proof of the main result (24) in this section to obtain: 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 the infinite sets of alternatives based on discrete numeral scales that have been proposed in the literature, for instance in Enguehard 2018. 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. At any rate, much of the discussion here is going to focus on cases where the set of worlds is finite to begin with.

One case where the assumption of finiteness will not be easily dispensable is certain results of Appendix A involving Minimal-World exhaustification.

[22]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.

4. A **does not contain several equivalent propositions.** While not crucial, this assumption simplifies the discussion in various places, and makes the next assumption viable. It lets us identify sentences and their denotations and omit the double brackets when they are too cumbersome.

5. **No two messages can have the same finite utility.** This assumption is sufficient (together with the assumption that A is finite) to make our use of arg max well-defined. One reason this is reasonable to assume is that if there are no equivalent propositions (as per the previous condition), and 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. For a given W and A, this assumption constrains the range of possible priors.

6. **All choices of P_0 compatible with the above are possible.** We have just restricted the choice of P_0, so we have to make explicit that we only do so to the degree that is necessary for our definitions to work. If arbitrary restrictions on P_0 are allowed, stability will be a trivial notion. **We will assume that valid choices are a dense subset of the domain of valid probability measures.**[23]

7. A **covers W: there is at least one true message at every world.** This serves simply to eliminate uninteresting corner cases, and can be easily made true by adding a tautology to A. When we discuss stability under negation, we will also implicitly assume that there is at least one false message at every world (which can similarly be made true by adding a contradiction to A).

2.2 Definition and non-probabilistic characterization of stability

We can now define stability. Stability is a property of the set of messages A: A is said to be stable when S_1 does not depend on P_0. In other words, at every world w, if for a certain setting of P_0 we have $S_1(w; P_0) = u^*$ for a certain u^*, then any setting of P_0 is such that $S_1(w; P_0) = u^*$ (and then we can just write $S_1(w)$). The reason such a property can be seen to be desirable is that it guarantees speakers and listeners do not need to agree on P_0 to communicate effectively. We will come back to whether stability is desirable or not in Section 4.2.

As a simple example, consider a case where you have two logically independent, overlapping propositions p_1 and p_2, and $A = \{p_1, p_2\}$. Recall that speakers have to use the least likely *a priori* among the true messages. Because neither proposition entails the other, neither is consistently more likely than the other, as we have proved above. Then, at a world w where they are both true, which of p_1 or p_2 is used will depend on P_0. This means that A is not stable.

Bar-Lev and Katzir (2022a) also define stability under negation, as follows: A is stable under negation if and only if $\hat{A}$ is stable, where $\hat{A}$ is the set of negated messages of A defined in (19). Below, we will call a system that is stable and stable under negation a *bilaterally stable* system. One might wonder why we define two kinds of stability in this way, rather than simply assuming that the alternative set is closed under negation. The problem is that alternative sets that are closed under negation are almost never stable, or not in interesting

[23] I was unable to construct an example of when this latter assumption could be violated that is reasonably close to any linguistics proposal, even dropping the assumption that A is finite; I believe it would have to involve countably infinite sets of alternatives, for reasons of dimensionality.

ways. Thus, we need to look at what happens when negation is present separately. Furthermore, the case of indirect implicatures (section 1.4) arguably provides motivation for looking at negation separately when it comes to alternative sets.[24]

(19) $\hat{A} = \{\neg u \mid u \in A\}$
where: $[\![\neg u]\!] = W - [\![u]\!]$

The practical definition of stability used by Bar-Lev and Katzir is not exactly the one we have just defined: instead, they allow messages corresponding to atomic propositions (sentences without logical operators) to be optimal in unstable ways. The intuition is that when they are used, these messages do not in fact compete with the more complex non-atomic messages, in line with Fox and Katzir's (2011) theory of alternative generation, and thus the model based on the full set A is not applicable. It seems interesting for our purposes to consider both notions, and I will prefer the phrase *quasi-stability* for the more liberal one. Formal definitions of stability, quasi-stability and their counterparts "under negation" are offered below:

(20) **Stability:** A is stable if there is no choice of a world w and priors P_0 and P_0' such that $S_1(w; P_0) \neq S_1(w; P_0')$.

(21) **Quasi-stability:** assume there is a distinguished set $A_0 \subseteq A$ of "atomic" messages. A is quasi-stable if there is no choice of a world w and priors P_0 and P_0' such that $S_1(w; P_0) \neq S_1(w; P_0')$ and $S_1(w; P_0) \notin A_0$.

(22) **Under negation:** A is (quasi-)stable under negation if $\hat{A}$ is (quasi-)stable (cf. (19)). In the case of quasi-stability, we take the atomic subset of $\hat{A}$ to be $\hat{A}_0$, where A_0 is the atomic subset of A.

It is possible to redefine stability in a way that does not directly refer to a decision-theoretic model. For this we can make use of the notion of the true alternative set at a world $\mathbb{T}_A(w)$, which is defined in (23) together with a notion of false alternative set. The characterization of stability as a property of true sets, given in (24) again along with its negative counterpart, turns out to be very simple. It is in fact equivalent to Dayal's (1996) *strongest answer* condition, which she argues is a property of answer sets; here, we are applying this condition to the alternative sets involved in pragmatic reasoning. A slightly more complicated characterization of quasi-stability can also be obtained, as seen in (25).

(23) **True and false alternative sets:** the true, resp. false, alternative set of a world $w \in W$, written as $\mathbb{T}_A(w)$, resp. $\mathbb{F}_A(w)$, is the set of propositions in A that are true, resp. false, at w:[25]

$$\mathbb{T}_A(w) = \{p : p \in A, w \in p\}$$
$$\mathbb{F}_A(w) = \{p : p \in A, w \notin p\}$$

[24] Of course, one might argue that in special-casing negation, we are reintroducing a monotonicity bias of some kind, which will undermine our results as an explanation for monotonicity biases. Specifically, the form the bias takes is that our results, including those who do not refer to stability under negation, essentially only apply for alternative sets that are not closed under negation, as only those sets are stable. It seems to the author that this does not undermine the explanatory potential of stability, given that the definition of stability is motivated in independent ways.

[25] If A is seen as the answer set of a question, then $\mathbb{T}_A(w)$ is what Karttunen (1977) proposes to be the denotation of the question at w.

(24) **Characterization of stability in terms of true sets:**

a. A is stable if and only if every true set is contains a least (strongest) element under entailment.

b. A is stable under negation if and only if every false set contains a greatest (weakest) element under entailment.

(25) **Characterization of quasi-stability in terms of true sets:**

a. A is quasi-stable if and only if every true set contains a least (strongest) element under entailment, or is such that all its minimal elements are atoms.

b. A is quasi-stable under negation if and only if every false set contains a greatest (weakest) element under entailment, or is such that all its maximal elements are atoms.

Proof. We first prove that all minimal elements of $\mathbb{T}_A(w)$ can be optimal at w. Consider then 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 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.

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 quasi-stability follow straightforwardly, as well as the "under negation" versions if we simply apply negation in the right places. □

To conclude, the property of stability introduced by Bar-Lev and Katzir (2022a) in the context of probabilistic models can be given a simple characterization based on the structure of the set of alternatives, which in the case of stability proper is essentially Dayal's (1996) strongest answer condition. We will use this characterization to derive certain results relating stability and monotonicity in the rest of this paper.

2.3 The consequences of stability for exhaustification

If we accept that alternative sets and answer sets are the same sort of objects and should be subject to the same constraints, perhaps because the alternative sets involved in pragmatic reasoning are the answer set of a latent question, then the equivalence between stability and the strongest answer condition provides independent motivation for stability as a property of alternative sets that the linguistic system would strive to maintain, and as a potential explanatory principle.[26] However, the fact that the strongest answer condition,

[26] While the notion of Question Under Discussion (QUD) is indeed frequently used in the literature, following Roberts (1996), it is not usually assumed that the QUD determines the alternative set. Most commonly, the

and therefore also stability, should apply to alternative sets can also be motivated in a more direct way. This is because (as Fox (2020) already discusses) the strongest answer condition can be restated with reference to the behavior of exhaustification operators: it is essentially equivalent to the condition that exhaustification should partition logical space.

In the semantics-pragmatics literature, "exhaustification" refers to any algorithmic formalization of the derivation of scalar implicatures and related inferences. An exhaustification operator is a function usually called EXH that takes two arguments, a proposition φ, the utterance, and a set of alternatives A, and computes the enriched meaning of the utterance, which is the conjunction of the literal meaning of the utterance together with its implicatures. A simple definition for an exhaustification operator is given in (26); this operator is used in Krifka 1993 to model the meaning of exclusives like *only*. What it does is simply conjoin the utterance with the negation of all its non-weaker alternatives. The case of more sophisticated operators is discussed in Appendix A.

(26) **Exhaustification operator:** the exhaustification of a proposition φ with respect to an alternative set A is given by:

$$\text{EXH}_A(\varphi) = \varphi \wedge \bigwedge_{\psi \in \text{NW}_A(\varphi)} \neg\psi$$

where: $\text{NW}_A(\varphi) = \{\psi \in A : \varphi \not\models \psi\}$

Exhaustification takes individual propositions as arguments, but we can of course apply it pointwise to a set, so that we are going to write $\text{EXH}_A(A)$ etc.

Exhaustification operators have also been used in the context of question semantics since Groenendijk and Stokhof 1984; they let us transform a set of overlapping propositions into a partition of the space of possible worlds. This is useful because natural approaches to the composition of questions will tend to yield a set of overlapping propositions, but a partition is arguably a better representation of the answerhood conditions of mention-all questions.[27] It is in fact for similar reasons that Dayal (1996) proposes that questions, analyzed as answer sets, are subject to a strongest answer condition: there should be exactly one strongest true answer at each world, a condition that we have seen is equivalent to stability.

The link between the ideas that questions should be represented by partitions, and that they should obey the strongest answer condition (a.k.a. stability), is not merely conceptual and can be given a formal form: these two desiderata are essentially equivalent. To state this precisely, we need to introduce two new notions: dominated alternatives, and the induced partition. Dominated alternatives, defined in (27), are alternatives that have their entire denotation covered by strictly stronger alternatives.

(27) **Domination:** a proposition p is *dominated* in A if for every world w where p is true, there is a proposition $q \in A$ that is true at w and strictly stronger than p, that is:

$$\forall w \in p.\, \exists q \in \mathbb{T}_A(w).\, q \subsetneq p$$

We can prove that dominated alternatives are exactly those that are never used by pragmatic speakers in our decision-theoretic model. Furthermore, it is straightforward to verify that dominated alternatives are mapped to contradictions by EXH as it is defined in (26).

QUD is seen as a partition of logical space which defines a notion of relevance. This restricts the alternative set, but its precise composition will also depend on the syntactic form of the utterance.

[27] Groenendijk and Stokhof (1984) actually apply exhaustification to a predicate rather than to a collection of propositions but the difference does not matter for our purposes.

Thus, we are going to ignore dominated alternatives, and focus on the behavior of the non-dominated subset.

(28) **Pragmatic speakers use non-dominated messages:** within the pragmatic model described in Section 2, the following two statements are equivalent:

a. Message φ is not dominated in A.

b. There exists a choice of prior P_0 and a world w such that $S_1(w; P_O) = \varphi$, that is, such that φ is the optimal message for a pragmatic speaker to use at w.

Proof. This follows from the proof of (24), and in particular the fact that the alternatives that are used are exactly the minimal elements of the true sets. □

The induced partition of the set of alternatives is defined in (29). This partition is defined in terms of the true sets: two worlds are in the same cell if their true sets are the same. The intuition is that this represents the maximal degree of informativity achievable by speakers using the alternative set: a speaker cannot be any more precise than specifying, for all alternatives, whether they are true or false. The induced partition will not change if we add logical functions of existing messages to the set of messages or if we apply interpretation rules that perform logical operations. This applies in particular to exhaustification: the exhaustified messages induce either the same partition as the original set of messages, or a strictly coarser partition. The only way to increase the granularity of the induced partition is to add new atomic concepts.

(29) **Induced partition:** the induced partition of a set of alternatives A is the collection of sets obtained by taking the inverse image of the true set operator. In other words, it is the partition obtained by taking the equivalence classes of the equivalence relation over worlds $\sim$ defined by:

$w \sim w'$ iff $\mathbb{T}_A(w) = \mathbb{T}_A(w')$

We can now state the main result of this section: stability, i.e. the strongest answer condition, is equivalent to exhaustification as defined in (30) producing a partition. The proof is fairly straightforward and follows from the fact that "φ is the strongest element in $\mathbb{T}_A(w)$" can be equivalently restated as "$\mathbb{F}_A(w) = \text{NW}_A(\varphi)$."[28]

(30) **Stability is equivalent to partition through exhaustification:** let A^* be the set of non-dominated alternatives within A. The following two statements are equivalent:

a. A is stable.

b. $\text{EXH}_A(A^*)$ is a partition of W.

When these statements are true, $\text{EXH}_A(A^*)$ is in fact the induced partition.

Proof. Suppose that $\text{EXH}_A(A^*)$ is a partition. Let w be a world and φ the non-dominated proposition such that $\text{EXH}_A(\varphi)$ is true at w. Since all elements of $\text{NW}_A(\varphi)$ are false at w, φ is the strongest true proposition at w. Generalizing to all worlds, A is stable.

[28] In Appendix A, we extend this result to two more sophisticated definitions of EXH, the IE operator (Fox 2007) and the MW operator (Schulz and van Rooij 2006; Spector 2006). The extended form of the result is slightly weaker, as the fact that the resulting partition is the induced partition needs to be specified rather than falling out on its own, and the assumption that A is finite is used in the proof. We also note that the result does not extend to the IE-II operator (Bar-Lev and Fox 2020). Unlike the earlier ones, this operator can create new partition cells out of dominated propositions.

Suppose A is stable. 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, 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 $\text{EXH}_A(\varphi)$ 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 $\text{EXH}_A(\varphi)$ Finally, if $\psi \in \text{NW}_A(w)$, then $\text{EXH}_A(\varphi)$ is false at w'. Thus $\text{EXH}_A(\varphi)$ is true at w' if and only if $w' \sim w$. Thus every non-dominated proposition is exhaustified into a 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 $\text{EXH}_A(A^*)$ are disjoint. Finally, since A is stable, a strongest true proposition can be found at any world (and it will necessarily be non-dominated). We can conclude that $\text{EXH}_A(A^*)$ is the induced partiton. □

This result can be argued to provide extra motivation for stability as a desideratum and as an explanatory principle, as long as we accept that exhaustification producing a partition, or more specifically producing the induced partition, is desirable. A possible line of argument (which we will expand on in the conclusion) is that this property is essential to letting speakers combine low production effort and high precision. In general, from a given set of overlapping alternatives, the maximal degree of precision that can be attained is if we specify the truth and falsity of every single alternative, or in other words we indicate the cell the of the induced partition that we are in. However, listing out all the true and false alternatives is a lot of effort on speakers' part. If exhaustification works out in such a way that each alternative is mapped to a cell of the induced partition, speakers can be as informative as they could ever hope to be, all while only actually uttering one simple sentence.

Before we move on, we have to mention that Fox (2020) proposes exactly this constraint for answer sets: they should be mapped by exhaustification to the induced partition — he calls this principle Question-Partition Matching (QPM). Fox argues on empirical grounds that QPM is superior to the strongest answer condition. This may appear to contradict our result. Part of the explanation is that Fox adopts a "Non-Vacuity" constraint, to the effect that all alternatives in the initial set should correspond to a partition cell; in contrast, we simply ignore what happens to dominated alternatives. The more fundamental difference is that Fox uses a variant of the IE-II operator, for which our result is false; cf. footnote 28 and Appendix A.

3 Stability and monotonicity

Having seen what stability is and why it is desirable, we still have to explain how it relates to the upward monotonic bias. In this section, we are going to show that in a variety of cases, stability constraints should lead us to adopt a lexicon of upward monotonic operators.

The pragmatic model presented in Section 2 does not make explicit reference to the lexicon: the alternatives are simply an arbitrary set of full sentences. In concrete situations, the alternatives in the alternative set are in fact related to one another, as their build-up includes common lexical elements. Thus the alternatives are determined in part by what is in the lexicon. Additionally, recall that we have not included a cost term in our utility function. There has to be a way in which our model incorporates the maxim of Manner.

In order both to account for the effect of Manner, and to model how the lexicon determines the alternatives, we are going to assume that the alternatives are obtained by applying "simple" logical operators to a set of atomic sentences, where "simple" means "up to a certain level of complexity": if the atomic sentences are $p_1, \dots, p_n$, then the alternatives have the form $F(\vec{p})$ where F is an n-place logical function, and all functions expressible within a certain complexity budget are included in the alternatives. The intuition (which we will come back to with critical eyes in Section 4.2) is that the atoms represent propositions like "John came", "Mary came", etc., and the other propositions are of the form "Q came" where Q is a quantifier phrase up to a certain, limited amount of complexity. We remain vague as to what level of complexity we are considering: single words, or non-coordinated structures, or something else; what matters is that we are considering a set of expressions that are often in competition with one another for the purposes of pragmatic reasoning.[29] Under these assumptions, the content of A will depend on the logical lexicon, given that the representation of a certain Boolean function can vary a lot in complexity depending on what primitives are available. It will then follow that if a certain property like stability is enforced on alternative sets, it will also constrain the logical lexicon. Note that in this setting, the partition induced by A is necessarily the same as the partition induced by the atoms; to simplify the notation, we will assimilate cells of this partition to individual worlds.

The bulk of this section is devoted to the exposition of formal assumptions and results, interleaved with proofs. A brief, high-level summary of the results is offered in Section 4.1 before we discuss the relevance of those results to broader questions.

3.1 The stable lexica of binary connectives

Bar-Lev and Katzir (2022a) explore the consequences of (quasi-)stability requirements on the lexicon of binary propositional connectives. The basic set-up here is that we have two atomic propositions p_1 and p_2; the set of worlds W is the 4-element set corresponding to all possible values of the vector $p_1 p_2$, that is, $W = \{00, 01, 10, 11\}$. Elements of the alternative sets A other than p_1 and p_2 are of the form $p_1 \star p_2$ where $\star$ is a 2-place Boolean function, assumed to correspond to a non-trivial lexicalized binary connective. Out of 16 possible binary functions, there are 10 that correspond to non-trivial connectives; the rest is made up of 2 projections (p_1 and p_2), their negations, and 2 constant functions ($\top$ and $\bot$).

The assumptions Bar-Lev and Katzir make go as follows:

(31) a. The two atomic propositions p_1 and p_2 (corresponding to the left-projection and right-projection in terms of connectives) are necessarily part of A.

b. Whether the two projections are stable as optimal messages does not matter when it comes to stability constraints: the messages p_1 and p_2 are allowed to be optimal only some of the time. Thus, we are evaluating the quasi-stability of A in the sense defined above.

c. All elements of A other than the projections correspond to symmetric (i.e. commutative) connectives. This brings the number of potential non-trivial connectives down to 6 ($\wedge$, $\vee$, $\oplus$, $\leftrightarrow$, NAND, NOR).

[29] We will also remain vague as to the precise source of alternative sets, but what we are saying here is broadly compatible with the theory of Fox and Katzir (2011). It is generally assumed that alternatives are sentences (syntactic objects) and not propositions (semantic objects) but this distinction does not matter to us and we assimilate both kinds of objects, so that our argument is compatible with both views; in particular, the lexicon we are talking about could be that of a language of thought.

Based on these assumptions, Bar-Lev and Katzir show through exhaustive search that no quasi-stable lexicons (where a lexicon is (quasi-)stable if it lets us build a (quasi-)stable alternative set) include any other connectives that $\wedge$, $\vee$ and NOR, and furthermore no bilaterally quasi-stable lexicons include any other connectives that $\wedge$ or $\vee$. Thus, bilateral quasi-stability selects for the lexicalization of upwards monotonic connectives within the symmetric connectives.

We can apply our characterization of quasi-stability to recover this result. Assume for instance that $\oplus$ is in the lexicon, so that $p_1 \oplus p_2$ is in A. At world 10, both p_1 and $p_1 \oplus p_2$ are true. $p_1 \oplus p_2$ is logically independent from p_1, so if we want quasi-stability, we need $p_1 \oplus p_2$ to be weaker than some other message that is true at 10, so that it will not be co-minimal with p_1. Because $p_1 \oplus p_2$ is true at only two worlds, 10 and 01, there is only one potential message that can be strictly stronger and true at 10, which is the message that is only true at 10, $p_1 \wedge \neg p_2$. However, this message corresponds to a non-symmetric connective ($\nrightarrow$, negated right-implication), and cannot be part of A. It follows that if A is quasi-stable, then $\oplus$ must not be lexicalized. With similar arguments, we can also show that NAND must not be lexicalized for A to be quasi-stable (looking at world 10 again, we find that we need $\oplus$), and that for A to be quasi-stable under negation, NOR and $\leftrightarrow$ cannot be lexicalized (looking again at world 10, NOR requires $\leftrightarrow$, and $\leftrightarrow$ requires the non-symmetric $\rightarrow$). It follows that, as far as symmetric connectives go, only subsets of $\{\wedge, \vee\}$ can be lexicalized to achieve bilateral quasi-stability, and one can verify that in fact they are all possible (adding a tautology and a contradiction to A to satisfy the covering requirements).

Our characterization also leads straightforwardly to a variant of this result, in terms of actual stability, which drops the symmetry requirements and therefore can be seen as more general. What we find is that upwards monotonic connectives are always needed to achieve bilateral stability, regardless of what other connectives are present.

(32) **Stability selects for upward-monotonic 2-place functions:**

a. If A is stable, the lexicon of binary connectives includes $\wedge$.
b. If A is stable under negation, then the lexicon of binary connectives includes $\vee$.
c. The only lexicon of binary connectives such that A is bilaterally stable and p_1 and p_2 are used as messages (both unembedded and under negation) is $\{\wedge, \vee\}$.

Proof. If A is stable, by (24), there is a strongest proposition p^* in $\mathbb{T}(11)$. p^* is stronger than both p_1 and p_2, since they are both in $\mathbb{T}(11)$. It follows that p^* can only be $p_1 \wedge p_2$.

Similarly, if A is stable under negation, there is a weakest proposition p_* in $\mathbb{F}(00)$. p_* must be weaker than both p_1 and p_2, so it can only be $p_1 \vee p_2$.

If A is bilaterally stable and p_1 and p_2 are used as messages, then we know by the above that $p_1 \wedge p_2$ and $p_1 \vee p_2$ are in A. Furthermore, p_1 is the strongest true message at some world (since it is used unembedded). p_1 cannot be the strongest true message at 11, because $p_1 \wedge p_2$, which is stronger, is true at 11. Thus p_1 is the strongest true message at 10. Then, other messages true at 10 and not weaker than p_1 are not in A: this includes $p_1 \oplus p_2$, $p_1 \nrightarrow p_2$, $\neg p_2$, and p_1 NAND p_2. Looking at 01, where the strongest true message must be p_2, we similarly get that $p_1 \nleftarrow p_2$ and $\neg p_1$ are not in A. Applying the same reasoning under negation, all messages false at 10 or 01 that are not stronger than p_2 or p_1 respectively are not in A: this includes $p_1 \leftrightarrow p_2$, $p_1 \leftarrow p_2$, $p_1 \rightarrow p_2$, and p_1 NOR p_2, as well as $\neg p_1$ and $\neg p_2$ which we have already ruled out. In conclusion $A = \{p_1, p_2, p_1 \wedge p_2, p_1 \vee p_2, \top, \bot\}$ (the tautology and contradiction are necessary to satisfy covering conditions) and the non-trivial connectives in the lexicon are exactly $\wedge$ and $\vee$. □

From the proof, we can already see how a stability constraint predicts the frequent lexicalization of upward-monotonic connectives: stability requires alternatives that are in certain entailment relationships to the atoms, and only upward-monotonic functions have the appropriate properties. Below, we derive similar results for n-place Boolean functions with arbitrary n.

3.2 The stable lexica of symmetric n-place functions

Let us then move on to the case of n-place logical functions: we now have n atoms $p_1, \dots, p_n$, and W is a 2^n-element set in bijective relation with the set of possible valuations of the vector $p_1 \dots p_n$. For any some subset I of $[\![1, n]\!]$, we will denote as w_I the world where the atoms whose index is in I are true, and all other atoms are false. As before, the alternative set A contains all atoms as well as elements of the form $F(p_1, \dots, p_n)$ where F is a "lexicalized" logical operator.

In this formal setting, 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. We can then speak not just of a monotonic function but also of a "monotonic proposition" (or a "convex proposition" when this notion will come up). 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. In the below, we will sometimes use phrases like "monotonic proposition"; one should keep in mind that we are really talking about logical functions.[30]

Moving back to the main topic, we want to characterize the constraints stability sets on the set of lexicalized n-place logical functions. A first idea, extending Bar-Lev and Katzir's (2022a) approach of the binary case, is to restrict our attention to symmetric functions, as defined in (33). These are the functions that do not distinguish between the different atoms. Note that all one-word quantifiers in English (not counting proper nouns and definite descriptions as quantifiers), such as *everybody* or *somebody*, are symmetric in this sense.

(33) **Symmetry:** an n-place boolean function f is symmetric if and only if for any permutation σ of size n, and for any vector $\vec{p} = p_1 \dots p_n$, we have:

$$f(p_{\sigma(1)}, \dots, p_{\sigma(n)}) = f(p_1, \dots, p_n)$$

There are 2^{n+1} symmetric boolean functions, which are in a canonical bijective relation to subsets of the integer interval $[\![0, n]\!]$. Indeed, such a function can only "count" the number of true propositions within the input, and it is characterized by the set of counts that it accepts.

We can now state our main result over symmetric functions:

(34) **Stability selects for *all* and *some* within the symmetric functions:** assume A includes only propositions corresponding to symmetric functions, other than the atoms. Then, if A is bilaterally quasi-stable, A does not include any other messages

[30] In less artificial settings, it is not as clear how to order worlds. Nevertheless, some theories require such an order to exist: an order on worlds is necessary to define the minimal-world exhaustification operator (Schulz and van Rooij 2006; Spector 2006) and to define convexity on propositions (Enguehard and Chemla 2019). In both cases, authors propose to derive the order from the true sets relative to a certain alternative set. This construction works precisely because the alternative set contains only propositions formed on upward-monotonic expressions. We describe the workings of the minimal-world operator in Appendix A.

than the atoms, and the propositions corresponding to the two constant functions, the existential quantifier, and the universal quantifier.

Proof. We start by assuming only that A is quasi-stable 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 $K \subseteq [\![0, n]\!]$ 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 $1 \leq k' \leq n$. In this case, φ_f is independent from any atom. Let I be a subset of $[\![1, n]\!]$ 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 φ' which is either φ_f or something stronger. Because φ_f is independent from the atoms, φ' is not an atom, and because A is quasi-stable, 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 $K \supseteq [\![1, n]\!]$, that is, f is a constant true function or *some*, or $K \subseteq \{0, n\}$, that is, f is a constant false function, *all*, *none*, or *all or none*.

Assume now that A is bilaterally quasi-stable: essentially the same reasoning lets us rule out any functions such that $k \notin K$ and $k' \in K$ can be found with $1 \leq k \leq n-1$ and $0 \leq k' \leq n-1$. This rules out *none* and *all or none*, leaving us with only the constant functions and the two upward-monotonic quantifiers. □

This result is clearly in the desired direction, since it picks out the two monotone increasing quantifiers of the Square of Opposition, which are known to be the most systematically lexicalized quantifiers across languages. Moreover, it corresponds to the set of true quantifiers in English (*everybody* and *somebody*, with the added *nobody* which is ruled out only by quasi-stability under negation). However, this results does not account for the distribution of quantifying determiners: there exist a number of lexical symmetric quantifying determiners in English and other languages that the bilateral stability constraint rules out, such as *many*, *most*, *two* or *few*. Note also that the reason we have to work with quasi-stability and not stability here is that with such a restricted lexicon, stability is not achievable: in the "middle", that is, in worlds where several atoms are true but not all of them, atoms are in unstable competition and there is no licit complex sentence that we could use in their stead.

3.3 The stable lexica of projectively symmetric n-place functions

In order to find larger stable lexicons, we can try to relax the symmetry constraint. A natural possibility is to allow domain restriction, which also makes our model arguably more realistic: natural language quantifiers have a restrictor and not just a scope, so that one may predicate them upon a subset of the overall domain. We will thus extend our attention to functions that are symmetric up to a domain restriction, which we will call *projectively symmetric* functions. The precise definition is given in (35).

(35) **Projective symmetry:** an n-place boolean function f is projectively symmetric if and only if, for a certain $k \leq n$, there is a sequence of k elements $i_1, \dots, i_k$ in $[\![1, n]\!]$ and a k-place symmetric function f' such that for all $\vec{p} = p_1 \dots p_n$:

$$f(\vec{p}) = f'(p_{i_1} \dots p_{i_k})$$

In particular, conjunctions and disjunctions of atomic propositions are projectively symmetric (they are universal or existential quantifiers applied to a subdomain), as well as their negations. This includes atoms and negated atoms. Unlike symmetric functions, projectively symmetric functions are not closed under conjunction and disjunction. They are, however, closed under negation.

We can now prove that the only projectively symmetric functions that can be stable in a lexicon including atomic propositions are upward-monotonic. Furthermore, stability requires that the lexicon should include all conjunctions and disjunctions of atomic propositions. The proof conceptually involves the following lemma (conceptually because we only use the easy direction), which is interesting enough that we state it separately: upward-monotonic functions are exactly those that can be written as the disjunction of a set of conjunctions of atoms.[31]

(36) **DNF characterization of monotonicity:** a function F is upward-monotonic if and only if $F(\vec{p})$ is equivalent to the disjunction of a set of conjunctions of atoms.

Proof. It is clear that a disjunction of conjunctions of atoms is upward-monotonic. Consider now an arbitrary upward-monotonic function f, and the associated proposition $\varphi_f = f(\vec{p})$. Recall that to each subset I of $[\![1, n]\!]$ we can associate the world w_I where exactly the atoms in I are true. We also adopt the notation $p_I = \bigwedge_{i \in I} p_i$. We can then define $\mathcal{I}_f = \{I : w_I \in \varphi_f\}$ and $\psi_f = \bigvee_{I \in \mathcal{I}_f} p_I$. If φ_f is true at a world w_I, then p_I is a disjunct of ψ_f, so that ψ_f is true at w_I. Thus φ_f entails ψ_f. Conversely, if ψ_f is true at w_I, then at least one disjunct of ψ_f is true at w_I, which means that there is $J \in \mathcal{I}_f$ such that p_J is true at w_I. This can only be the case if $J \subseteq I$, which means that $w_J \leq w_I$. φ_f is true at w_J and f is upward-monotonic, so φ_f is true at w_I. Then, ψ_f entails φ_f, and in conclusion φ_f and ψ_f are equivalent, which means that φ_f is equivalent to a disjunction of conjunctions of atoms. □

(37) **Stability selects for upward-monotonic projectively-symmetric functions:** assume that A includes the atoms and that it includes only propositions corresponding to projectively symmetric functions. Then:

a. If A is quasi-stable, all propositions in A that are false at the world where all atoms are false correspond to upward-monotonic functions.
b. If A is stable, A includes all non-empty conjunctions of atoms.
c. If A is bilaterally quasi-stable, A only includes propositions corresponding to upward-monotonic functions.
d. If A is bilaterally stable, A includes all conjunctions and disjunctions of atoms.

Proof. As a lemma, let us first prove that a projectively symmetric proposition non-trivially entailing an atom is a conjunction of atoms. Let φ be a projectively symmetric proposition such that $\varphi \vDash p_i$ for some i. We can find $J \subseteq [\![1, n]\!]$ (the domain restriction of φ) and $K \subseteq [\![0, |J|]\!]$ (the accepted counts of φ) such that for arbitrary I, φ 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 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 p_J, the conjunction of all atoms whose index is in J, or $K = \varnothing$, and φ is a contradiction. Note that

[31] All Boolean functions can be written as a disjunction of conjunctions of possibly-negated atoms; this is their Disjunctive Normal Form (DNF). A function is upward-monotonic if and only if it admits a DNF without negation. A similar result obtains for a function's Conjunctive Normal Form (CNF), which is a conjunction of disjunctions of possibly-negated atoms.

in the former case, necessarily $i \in J$.

Taking the negation (recall that projective symmetry is closed under negation) and replacing atoms by their negation in the proof, we also obtain the following result: if φ is projectively symmetric and $p_i \vDash \varphi$, then φ is a disjunction of atoms (including p_i) or a tautology.

Now, 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 p_J for some J such that $I \subseteq J$ by the above lemma; furthermore since φ^* is true at w_I, φ^* is in fact p_I.

If A is stable, there is a strongest proposition at w_I for all I. By the above, that proposition is p_I as long as $I \neq \emptyset$. Thus A includes all p_I's for non-empty I. This proves (b).

If A is quasi-stable, consider an arbitrary non-atomic proposition φ in A. Suppose φ is false at $w_\emptyset$. As before, let $\mathcal{I}_\varphi = \{I : w_I \in \varphi\}$, and define $\psi = \bigvee_{I \in \mathcal{I}_\varphi} p_I$. If φ is true at w_I for some I, then p_I is a disjunct of ψ and ψ is true at w_I as well, so that $\varphi \vDash \psi$. Furthermore, for $I \in \mathcal{I}_\varphi$, because A is quasi-stable, either there is a single strongest proposition in $\mathbb{T}_A(w_I)$, or the minimal elements of $\mathbb{T}_A(w_I)$ are atoms. In the former case, the single strongest proposition is p_I by the above. Because φ is in $\mathbb{T}_A(w_I)$, $p_I \vDash \varphi$. In the latter case, the minimal elements are in fact all the p_i's for $i \in I$ (since they are the only atoms in $\mathbb{T}_A(w_I)$, and they are all independent). By assumption, φ is weaker than at least one minimal element, so at least one p_i, and it follows that $p_I \vDash \varphi$. Then $p_I \vDash \varphi$ for any $I \in \mathcal{I}_\varphi$, which entails that $\psi \vDash \varphi$. In conclusion ψ and φ are equivalent and φ is a disjunction of atomic conjunctions, and therefore upward-monotonic. This proves (a).

If A is bilaterally quasi-stable, the earlier results, applied under negation, tell us that if there is a weakest false proposition at w_I with $I \subsetneq [\![1, n]\!]$, then that proposition is the disjunction of atoms not in I: $q_I = \bigvee_{i \notin I} p_i$. Otherwise, the atoms false at w_I are maximal in $\mathbb{F}_A(w_I)$. It follows that any proposition false at w_I is at least as strong as q_I, and therefore, is false at $w_\emptyset$. Then , any proposition true at $w_\emptyset$ is a tautology, and is therefore upward-monotonic. Together with (a), this proves (c).

If A is bilaterally stable, there is a single weakest false proposition at every world; by the above that proposition is q_I. Thus all disjunctions of atoms are in A. Together with (b), this proves (d). □

What these results show is that only upward-monotonic operators can be lexicalized, in order to obtain a bilaterally quasi-stable lexicon of projectively symmetric functions. Furthermore, we in fact need all the atomic conjunctions and disjunctions in order to achieve bilateral stability. While we omit the (straightforward) proof here, this is also the case if we forget about projective symmetry and start from the assumption that operators have to be upward-monotonic (or just monotonic). Either way, stability pushes us towards a system where we always use the largest true conjunction of atoms, or, under negation, the largest false disjunction of atoms: essentially, we pinpoint the world by listing the positive or negative examples of what we are talking about (e.g. listing the people who came) and implying that the possibilities we did not mention are on the other side. What is remarkable here is that we are not using at all the common intuition that this pattern is motivated by considerations of efficiency; instead it follows entirely from relations of logical entailment between propositions.

It is easy to verify that any other upward-monotonic operators (such as some version of *most* or *at least two*), including also non-projectively symmetric ones (such as *Mary or both boys*) can be added without endangering stability; these extra alternatives will simply never be used by S_1. This is because, respectively, the strongest / weakest possible upward-

monotonic proposition that is true / false at a given world is an atomic conjunction / disjunction. One may of course find this property of the model here unpleasant: speakers do in fact use *most*.[32]

3.4 The stable lexica of arbitrary n-place functions?

It would be satisfying, of course, to drop the projective symmetry condition and prove that all stable lexicons are made of upward-monotonic operators. Let us immediately observe that this is false. The choice of A in (38), not based on monotone increasing operators, can be verified to be bilaterally stable. This set of messages simply includes all messages that are true at exactly one world, as well as all messages that are false at exactly one world. Messages like this are trivially optimal when they are true, since they are as informative as is possible, even before any pragmatic reasoning. These messages are neither projectively symmetric nor monotonic (except for 4 of them that are essentially *some*, *none*, *all* and *not all*, the corners of the Square of Opposition). The atomic propositions are not part of the set, but we can add them without endangering stability; they will simply never be used. All this is, of course, very much unlike anything observed in natural language.

(38) $A = X \cup \{\neg\varphi : \varphi \in X\}$
where: $X = \{p_I \wedge \neg q_I : I \subseteq [\![1, n]\!]\}$
$$= \left\{\left(\bigwedge\nolimits_{i\in I} p_i\right) \wedge \left(\bigwedge\nolimits_{i\notin I} \neg p_i\right) : I \subseteq [\![1, n]\!]\right\}$$

The set of messages in (38) is not related by inclusion to the set of all atomic conjunctions and disjunctions, so that it is also not true that the atomic conjunctions and disjunctions are the smallest possible stable set (neither is it true of the set in (38)).[33] The two sets are also similar in size: the size of this lexicon is 2^{n+1} (or $2^{n+1}+n$ if we add the atoms), while the set of atomic conjunctions and disjunctions is only slightly smaller at $2^{n+1} - n$.

What we would want then is a constraint that rules out the set in (38) and similar examples, and that is not some form of symmetry. In the $n = 2$ case, we have used the constraint that all atoms should actually be used, but this is not sufficient for arbitrary n. One thing we can remark about the optimal propositions under negation in (38) is that they are not convex in the sense defined in Section 1: they have "holes" in their truth set. Convexity is argued by Enguehard and Chemla (2019) to be active in determining the distribution of embedded exhaustification, perhaps because propositions verifying it are easier to think about. Thus, it would be natural to place the constraint that the lexicon should be restricted to convex functions. Unfortunately, this does not suffice to make stability entail upward-monotonicity. Instead, we can derive a slightly different result, stated in (39): stable alternative sets must to be built with upward monotonic operators if we restrict our attention to

[32] As Bar-Lev and Katzir (2022a) point out, the binary model already predicts that disjunction will never be used other than under negation, which is arguably unproblematic in their setting as we know that unembedded disjunction is only used in situations of ignorance, and we are ignoring those situations here. This argument will not work for *most*, which does not generally convey ignorance, and is usually used unembedded.

[33] Note that for $n = 2$, the set of messages in (38) corresponds to the 8-element lexicon $\{\wedge, \nrightarrow, \nleftarrow, \text{NOR}, \vee, \rightarrow, \leftarrow, \text{NAND}\}$. It is therefore a superset of the set of atomic conjunctions and disjunctions that are neither constants nor atoms, which corresponds to the lexicon $\{\wedge, \vee\}$. This is why the bilateral stability of this set does not contradict our earlier result that all bilaterally stable lexicons build up on $\{\wedge, \vee\}$. The inclusion stops being true for $n \geq 3$.

functions whose negation is convex, rather than if we look directly at convex functions.[34]

(39) **Stability selects for upward-monotonic convex-under-negation functions:** assume that A includes the atoms and that is includes only propositions corresponding to functions whose negation is convex. Then:

a. If A is quasi-stable, all propositions in A that are false at the world where all atoms are false correspond to upward-monotonic functions.
b. If A is stable, A includes all non-empty conjunctions of atoms.
c. If A is bilaterally quasi-stable, A only includes propositions corresponding to upward-monotonic functions.
d. If A is bilaterally stable, A includes all conjunctions and disjunctions of atoms.

Proof. We will just show that if a proposition whose negation is convex is the single minimal element of $\mathbb{T}_A(w_I)$ for $I \neq \emptyset$ then it is the conjunction p_I of the atoms true at I. The rest of the proof is as for (37).

Suppose then that φ^* is the strongest element of $\mathbb{T}_A(w_I)$ for some world w_I where the atoms that are true have their indices in $I \neq \emptyset$. Because φ^* entails all the p_i's with $i \in I$, φ^* is false at $w_\emptyset$. Now take J a strict superset of I. We have $w_\emptyset \leq w_I \leq w_J$: w_I is in-between w_J and $w_\emptyset$. Because the negation of φ^* is convex, and φ^* is true at the in-between world in this chain, φ^* cannot be false at both extreme points w_J and $w_\emptyset$. Therefore, φ^* is true at w_J. This holds for any $J \supsetneq I$ and therefore φ^* is equivalent to p_I. □

The reason this result is unfortunate is that there is no independent motivation for a preferrence for functions that are convex under negation: as Enguehard and Chemla (2019) discuss, the enriched meaning of sentences containing scalar items, taking implicatures into account, is generally a proposition that is convex but not monotonic, and whose negation is not convex. Still, it shows that stability as a goal can lead to a preferrence for upward-monotonic operators, also with other constraints on the lexicon than symmetry.

3.5 The stable generative lexica

The final result relating stability and monotonicity that we will offer is based on a more structured perspective on the logical lexicon. In the above, we have considered that any set of logical functions where each individual function satisfies the constraints we have placed is up for lexicalization. We also have considered each number of atoms separately. In natural language, of course, the same logical words are used regardless of the size of the domain, at least above three. We would therefore like to derive results applying to collections of logical functions applying to arbitrary arities. These collections should furthermore have some structure, such that the possible functions for different arities are related.

One approach we can adopt is to assume that the set of expressible logical functions is obtained generatively, through composition of a small number of primitives. We will assume that we have a set B of logical primitives. The set of alternatives is then identified to the set of sentences obtained by applying the functions in B an arbitrary number of times to the atoms, as described in (40) (with the obvious semantics). To get the alternatives for a domain of size n, we simply restrict our attention to formulas where no atom of index

[34] I think one can obtain a somewhat messy result with unembedded convexity: if A is bilaterally stable and the lexicon is made of convex functions, and if the atoms are used unembedded and under negation, then the restriction of messages in the alternative set to the set of worlds where some but not all of the atoms are true is upward-monotonic.

higher than n is present. In this construction the atoms are to be understood as abstract variables and not as concrete propositions; their instantiation will vary with the topic.

(40) **Generative model of the alternatives:**

a. For any i, p_i is an sentence.

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

As before, we can make our set of sentences correspond to a set of logical functions, by associating each sentence to the function which one needs to apply to the atoms to obtain the same semantics. For each arity n, we get a set of expressible functions of this arity, C_n, and we can put them together to form the overall set $C = \bigcup_n C_n$.[35] The generative model imposes a certain structure on C: it contains all the projections (corresponding to atomic sentences), and is closed under composition. Such a structure is called a clone, and B is called the basis of the clone. Formal definition of these notions are given in (41) and (42).

(41) **Clones:** let C be a collection of logical functions of varying arities, and C_n be the elements of C of arity n. C is a clone if:

a. All the projections are in C: for any n and any $i \leq n$, the function π^i_n such that $\pi^i_n(p_1, \dots, p_n) = p_i$ is in C_n.

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

(42) **Basis of a clone:** the set of logical functions B is said to be a basis for the clone C if C is the smallest clone including B.

The upward-monotonic functions form a clone M generated by the basis of binary connectives and constant functions $\{\wedge, \vee, 0, 1\}$. The set of all logical functions L is also a clone, generated for instance by $\{\wedge, \neg\}$. A result due to Post (1941)[36] is that M is a maximal proper subclone of L: there is no clone having M as a proper subset other than L. There are five such maximal subclones, including M. The other four are the truth-preserving functions P_1, the falsity-preserving functions P_0, the affine functions A and the self-dual functions D.[37]

As we have already seen, it is possible to achieve bilateral stability if all logical functions are allowed, using the propositions in (38). However, what we can show is that if there is any restriction to the set of expressible propositions (as one might argue there should be for reasons of economy), then that restriction should be to the upward-monotonic functions.

(43) **Stability selects for the upward-monotonic clone:** assume that for any n, the alternative set is obtained by taking the functions of the appropriate arity in a given

[35] The mapping between C and the fragment derived from the generative rules is not exactly bijective, because each formula will denote an infinity of functions of varying arities: if the variable with the highest index in a formula is p_m, then we can interpret the formula as a function of arity n for any $n \geq m$.

[36] 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).

[37] 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. We are going to use the fact that the affine functions and self-dual functions are always true at exactly half of their possible inputs, with the exception of the constants in the case of affine functions.

clone $C \subsetneq L$. Then:

a. If the alternative set is always quasi-stable, then $C \subseteq M$.
b. The alternative set is always bilaterally stable if and only if $C = M$.

Proof. Since C is a proper subset of L, C is a subset of at least one of the five maximal subclones. Because of our covering assumptions (at every world there is at least one true proposition and one false proposition), C is not a subset of P_0 or P_1. It must then be a subset of A, D or M.

Assume C is such that we always get quasi-stable alternative-sets and suppose C is a subset of D or A. Then, all non-constant functions in C are true at exactly half of their inputs. Let f be a non-constant element of C_n for some $n \geq 2$, and let I be a non-empty subset of $[\![1, n]\!]$ such that $f(\vec{p})$ is true at w_I (such a I exists since f is true for half its inputs). If $f(\vec{p})$ is the least element of $\mathbb{T}(w_I)$, then it is stronger than at least one atom. Since atoms are true at half of worlds, $f(\vec{p})$ is in fact equivalent to the atom. Thus f is a projection. If the least element of $\mathbb{T}(w_I)$ is $f'(\vec{p})$ for $f' \neq f$, then f is strictly weaker than f'. Since they are both true at half their inputs, this is impossible. If $\mathbb{T}(w_I)$ has no least element, its minimal elements are atoms, and f must be weaker than at least one projection, which again means it is equal to that projection. In conclusion, f is a projection, and generalizing, C 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 C since we would then have non-projections for $n \geq 2$. Then C contains only projections and constants, and is therefore a subset of M. We can then conclude that C is always a subset of M.

We have seen earlier that for $n = 2$, for the alternative set to be bilaterally stable, we need conjunction and disjunction. Furthermore, to satisfy the covering assumptions, we need a contradiction and a tautology. Then, if C is such that the alternative set is always bilaterally stable, C contains a full basis of M and therefore $C = M$. In the other direction, it is straightforward to verify that if $C = M$, the alternative set is always bilaterally stable: at every world we use the largest conjunction of atoms true at that world, and under negation we use the largest disjunction of atoms false at that world. □

Note that the full set of all upward-monotonic functions is larger than the atomic disjunctions and conjunctions that are strictly needed to achieve bilateral stability, as we have proved above; for instance, the function corresponding to the proposition $(p_1 \wedge p_2) \vee p_3$ is included (this function is not projectively symmetric). At any rate, once again, a stability constraint will single out the upward-monotonic logical functions.

4 Discussion

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

In Section 1 of this article, we have argued that the logical property of monotonicity, and especially upward-monotonicity, has some kind of distinguished position in natural language as well as possibly cognition. The logical lexicon is made chiefly of upward-monotonic operators, and includes a variety of items that mark the monotonicity status of syntactic contexts through their distribution. In several instances, theories of certain phenomena at the semantics-pragmatics interface, such as implicature derivation or question semantics,

need to stipulate an asymmetry between upward-monotonic propositions and other propositions, or some related distinction, to make correct predictions. Finally, there is some evidence that upward-monotonic logical functions are easier to reason about at the cognitive level. We call this pattern the upward-monotonic bias. Explanations for the bias — which have tended to focus on the lexicalization facts, also because the other phenomena can be argued to be consequences of them — have been offered in the literature, but they tend to rely on specific assumptions on conceptual primitives and therefore merely displace the question.

In Section 2, we have introduced Bar-Lev and Katzir's (2022a) notion of communicative stability, defined in the context of probabilistic models of pragmatics: stability is the property that the predicted behavior of the pragmatic speaker S_1 does not depend on the probabilistic prior. We have defined several formal operationalizations of this idea (stability proper or quasi-stability, unilateral or bilateral) and shown that they can be reformulated in formal-logical terms; specifically, they are equivalent to variants of Dayal's (1996) strongest answer condition applied to alternative sets. This condition is in turn equivalent to a requirement that exhaustification should turn the set of alternatives into the most fine-grained possible partition of logical space.

In Section 3, we have proved a series of results to the effect that, under various additional assumptions, if the alternatives are obtained by applying a certain set of "lexicalized" (that is, expressible in a not too complicated way) operators to atomic sentences, then one should only lexicalize upward-monotonic operators in order to obtain stable alternative sets. In general, Bar-Lev and Katzir's (2022a) proposal that stability can predict the upward-monotonic bias, made specifically in relation to binary connectives, can be extended to the whole logical lexicon.

Concretely, we have shown that stability selects for the upward-monotonic connectives, conjunction and disjunction, within all binary connectives; these are in fact the connectives that are commonly lexicalized. Stability also selects for the existential and universal quantifiers within all symmetric logical functions; these are the one-word quantifiers that are most commonly lexicalized, such as *everybody* and *somebody* in English. Within functions that are symmetric up to a domain restriction, stability demands again existential and universal quantifiers, and rules out non-upward-monotonic functions. One-word quantifying determiners are indeed most commonly upward-monotonic with respect to their second argument. Finally, if we assume that the lexicon is generated recursively from a small set of primitives, the only way to achieve stability without just lexicalizing all possible functions is to generate exactly the upward-monotonic functions. The deeper reason that all these results hold is that stability demands that alternatives should be in certain logical relations to simple sentences without logical operators, and since these sentences are upward-monotonic in a certain sense, more complex sentences have to be upward-monotonic as well.[38]

Each of our results relates some kind of preference for upward-monotonicity to some kind of stability constraint. The motivation for stability has to do with the behavior of pragmatic speakers in probabilistic models, or alternatively with the potential for implicature

[38]We might want to also explain why simple sentences are upward-monotonic: perhaps we can imagine a language where the primitive relation is "not be part of a set" rather than "be part of a set" (I naïvely assume here that the sets themselves, which correspond to content words, are not questionable). One possible explanation is that it is not clear we can build a collection of logical functions with appropriate closure properties out of this; recall that downward-monotonicity is not closed under composition. Another possible explanation, along the lines of Enguehard and Spector 2021, is that the simplest sentences in such a language would not be as informative as in our languages.

derivation, neither of which seem related to questions of monotonicity in principle. For this reason, it seems plausible in light of our results that Bar-Lev and Katzir's (2022a) communicative stability has a role to play in explainging the bias towards upward-monotonicity in natural language and cognition in a non-circular manner. This is all the more so that, as we have discussed in Section 1, there have been no other good candidates so far.

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 3.5. 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 languages 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 stability (or equivalently, Dayal's (1996) strongest answer condition), as shown in Section 2.3. To get stability, in turn, under plausible additional constraints on the lexicon, we need upward-monotonic operators only, as shown throughout Section 3. 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 Some challenges

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

The desirability of insentivity to priors The initial motivation we offered for stability is that it should enable speakers to communicate without agreeing on the prior. It is however not clear that this is desirable at all. Indeed, if stability 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 (44), 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 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.

(44) 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. This undermines the motivation for stability as an explanatory principle; while we have seen in Section 2.3 that stability could also be motivated without reference to probabilities, it is suspicious that some observed pragmatic phenomena are best analyzed using alternative sets that appear to violate it.

One possible way to reconcile 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 messages and the negated messages, while the choice of the specific message within a given, stable set is not sensitive to the prior. This would be consistent with examples like (44) and compatible with stability as an explanation for the upward-monotonic bias. In such a system, the speaker-oriented notion of stability that we have been using would not hold of the whole set of messages. 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 message, they know the speaker takes the prior to be in such and such way, and because we have stability within the negated messages, 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.[39]

Limits of the formal setting The relevance of our formal settings to lexicalization patterns is also worthy of criticism. Indeed, our choice of formal settings in Section 3 was primarily driven by mathematical convenience, and it is arguably a poor fit for actual linguistic situations in various ways. This is especially the case for $n \geq 3$. It is not clear that natural languages actually lexicalize any Boolean function of arity more than 2; instead, what we are trying to model is the use of quantifiers. Quantifiers actually operate on predicates; furthermore in many cases what we observe is quantifying determiners with two predicate arguments, a restrictor and a scope. What we have done above is that we have identified the scope predicate P to the sequence of atomic sentences it can form ($P(x)$), assuming all combinations are possible, and leaving 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 the internal logical structure the restrictor may have. Finally, we assume all alternatives have the same predicate, while all logically possible restrictors are alternatives, which is an odd asymmetry. All in all, what we have is a model of how stability affects the competition between quantifying words, but not really the competition between quantified statements in general. This can be seen in the fact that we predict that words like *most* will never be

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

used, since they are always less useful than a suitably restricted universal statement. This is of course incorrect and probably betrays the fact that our alternatives are wrong. It is nevertheless in the author's view a good first step, that later models of the impact of lexicalization choices on restrictors and scopes can build on.

A related objection 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,[40] More generally, the representation of quantifiers will tend to grow with domain size. The actual cognitive representation of quantifiers is very probably not dependent on domain size in any 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.

4.3 Conclusion

With the caveats discussed above, we have argued in this paper that a stability desideratum predicts the dominance of upward-monotonic logical functions in language, beyond the cases considered by Bar-Lev and Katzir (2022a), and that it also is necessary for the derivation of strong implicatures. Our final claim is that these two facts should be put together: 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. Furthermore, we have not explained everything that we have claimed is a reflex of the upward-monotonic bias, such as the existence of NPIs. Let us also note that there is no reason in principle that the upward-monotonic bias should have a single simple explanation. The explanation we offer here is not incompatible with other explanations that involve additional 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: stability and sophisticated exhaustification operators

A.1 Main result

In this appendix, we extend the result of Section 2.3 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"

[40] 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.

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 Section 2.3 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.

We retain all the formal assumptions of Section 2: to recapitulate the essential bits, W is a set of worlds, and A a finite set of propositions over those worlds that covers W (there is at least one true proposition at each world). We continue to identify sentences to propositions and propositions to sets of worlds, so that we generally omit double brackets and write indifferently "φ is true at w" and "$w \in \varphi$". For clarity, we will refer to the operator defined in Section 2.3, in (26), as EXH^0.

Let us then define the exhaustification operators. We start with the IE operator, which is defined in several steps.

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

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

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

(48) **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 $\text{EXH}^{IE}_A(\varphi)$ defined by:

$$\text{EXH}^{IE}_A(\varphi) = \varphi \cap \bigcap_{\psi \in \text{IE}_A(\varphi)} (W - \psi)$$

We then define the MW operator. This operator requires on order on worlds: such an order can be derived from the true sets, and is called the induced order.[41] 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.

(49) **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:

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

(50) **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 $\leq_A$:

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

All the notions defined above are defined with respect to the alternative set A. We have

[41] 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.

indicated this dependency as a subscript, but to keep notation simple, we will often leave the dependency on A implicit from now on.

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.

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

$$\text{IE}(\varphi) = \{\psi \in A : \text{EXH}^{MW}(\varphi) \vDash \neg\psi\}$$

(Spector 2016)

It follows in particular from (51) that $\text{EXH}^{MW}(\varphi)$ is always at least as strong as $\text{EXH}^{IE}(\varphi)$. Spector (2016) also shows that when the basic operator EXH^0 does not yield a contradiction, the other two operators give the same result:

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

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

The fact that A is finite lets us take for granted a variety of things. As before, it 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.[42]

Let us finally state the main result, which is that stability is equivalent to the fact that or MW exhaustification of the non-dominated messages gives us the induced partition.

(53) **Stability is equivalent to exhaustification producing a partition:** let A^* be the set of non-dominated propositions in A. The following three statements are equivalent:

a. A is stable.
b. $\text{EXH}^{MW}(A^*)$ is a partition of W and this partition is the same as the partition induced by A.
c. $\text{EXH}^{IE}(A^*)$ is a partition of W and this partition is the same as the partition induced by A.

Proof. From Spector's result (52) and our earlier result in (30), it follows immediately that if A is stable, then $\text{EXH}^{MW}(A^*)$ and $\text{EXH}^{IE}(A^*)$ are the induced partition. This proves that (a) entails (b) and (c).

Suppose now that $\text{EXH}_{IE}(A^*)$ is the induced partition. Let φ be a non-dominated proposition. 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 a partition cell. Since $\text{EXH}_{MW}(\varphi)$ is at most as strong as $\text{EXH}_{IE}(\varphi)$, which is a partition cell, then $\text{EXH}_{MW}(\varphi) = \text{EXH}_{IE}(\varphi)$. This proves that (c) entails (b).

Finally, suppose that $\text{EXH}_{MW}(A^*)$ is the induced partition. Let w be a world and φ the unique non-dominated proposition such that w is minimal for φ. If φ is true at some world w', there must be a minimal φ-world somewhere below w' (here again we use the fact that A is finite). Then by the starting assumption, that world is in the cell of w. Thus $w' \geq w$.

[42] In Spector's result (51), 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 example in Section A.2. The result in (52) does not depend on such an assumption.

Since φ is true anywhere above w by definition of the order, 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 $\varphi \vDash \psi$. Generalizing, at every world, the unique non-dominated proposition whose MW-exhaustification is true is also the strongest true proposition. Then A is stable, and (b) entails (a). □

A.2 Some counterexamples showing the necessity of our assumptions

The result in (53) is slightly weaker than the earlier one in Section 2.3: the fact that the partition is the induced partition has to be specified, rather than falling out on its own. This is because unlike EXH^0, the IE and MW operators can produce coarse partitions, and they do not do so in the same cases.

Here is a case with four worlds where the lexicon is not stable, but IE and MW exhaustification produce a coarse partition: take $p = \{w_1, w_2\}$, $p' = \{w_1', w_2'\}$, and $q = \{w_2, w_2'\}$. Notice that the choice between p and q in w_2 and between p' and q in w_2' depends on the prior. Nevertheless, we have $\text{EXH}^{MW/IE}(p) = \{w_1\}$, $\text{EXH}^{MW/IE}(p') = \{w_1'\}$ and $\text{EXH}^{MW/IE}(q) = q = \{w_2, w_2'\}$, so that $\text{EXH}^{MW/IE}(A)$ is a partition (for both operators). This partition only has three cells, compared to four in the original induced partition.

It is also not true that EXH^{IE} and EXH^{MW} output partitions in the same cases, even when talking about coarse partitions. In fact, both implications are false. 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 an unstable set of alternatives. 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. The alternative set is not stable because there is no strongest 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.[43]

[43] 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; I leave the determination of such a constraint to future work.

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 (54). 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.

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

$$\begin{aligned}\textsc{exh}_A^{IE-II}(\varphi) &= \{w : \mathbb{F}_A(w) = \mathrm{IE}_A(\varphi)\} \\ &= \left(\bigwedge\nolimits_{\psi \notin \mathrm{IE}(\varphi)} \psi\right) \wedge \left(\bigwedge\nolimits_{\psi \in \mathrm{IE}(\varphi)} \neg\psi\right)\end{aligned}$$

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 optimal proposition at w'. Yet, because $\mathrm{IE}(q) = \varnothing$, we have $\textsc{exh}^{IE-II}(q) = q \wedge p_1 \wedge p_2 = \{w'\}$, so that $\textsc{exh}^{IE-II}(A)$ is the induced partition.

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.

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.

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 (2013). *Logic in grammar: Polarity, free choice, and intervention*. Oxford University Press.

Crnič, Luka (2014a). "Against a dogma on NPI licensing". In: *The art and craft of semantics: A festschrift for Irene Heim*, pp. 117–145.

Crnič, Luka (2014b). "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.

Enguehard, Émile (2018). "Comparative modified numerals revisited: scalar implicatures, granularity and blindness to context". In: *Semantics and Linguistic Theory*. Ed. by Sireemas Maspong, Brynhildur Stefánsdóttir, Katherine Blake, and Forrest Davis. Vol. 28, pp. 21–39. DOI: 10.3765/salt.v28i0.4403.

Enguehard, Émile and Emmanuel Chemla (2019). "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 (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.

Geurts, Bart and Frans van Der Slik (2005). "Monotonicity and processing load". In: *Journal of semantics* 22.1, pp. 97–117. DOI: 10.1093/jos/ffh018.

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.

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

Hamblin, Charles L. (1976). "Questions in Montague English". In: *Montague grammar*. Ed. by Barbara H. Partee. Academic Press, pp. 247–259. DOI: 10.1016/B978-0-12-545850-4.50014-5.

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

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.

Jeretič, Paloma (2022). "Neg-raising as a scaleless implicature". In: *Semantics and Linguistic Theory*. Vol. 32. DOI: 10.3765/salt.v1i0.5367.

Kadmon, Nirit and Fred Landman (1993). "Any". In: *Linguistics and philosophy* 16.4, pp. 353–422.

Karttunen, Lauri (1977). "Syntax and semantics of questions". In: *Linguistics and Philosophy* 1, pp. 1–44.

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.
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.
Krifka, Manfred (1995). "The semantics and pragmatics of polarity items". In: *Linguistic Analysis* 25, pp. 209–258.
Kuhn, Jeremy (2022). "The dynamics of negative concord". In: *Linguistics and Philosophy* 45.1, pp. 153–198.
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.
Post, Emil L. (1941). *The Two-Valued Iterative Systems of Mathematical Logic*. Vol. 5. Annals of Mathematics studies. Princeton University Press.
Roberts, Craige (1996). "Information structure in discourse: towards an integrated formal theory of pragmatics". In: *OSU working papers in linguistics* 49, pp. 91–136.
Sauerland, Uli (2004). "Scalar implicatures in complex sentences". In: *Linguistics and philosophy* 27.3, pp. 367–391.
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.
Spector, Benjamin (2006). "Aspects de la pragmatique des opérateurs logiques". PhD thesis. Université Paris-7.
Spector, Benjamin (2008). "An unnoticed reading for *wh*-questions: elided answers and weak islands". In: *Linguistic Inquiry* 39.4, pp. 677–686.
Spector, Benjamin (2014). "Global positive polarity items and obligatory exhaustivity". In: *Semantics and Pragmatics* 7.11, pp. 1–61. DOI: 10.3765/sp.7.11.
Spector, Benjamin (2016). "Comparing exhaustivity operators". In: *Semantics and Pragmatics* 9, pp. 11–1.
Uegaki, Wataru (2022). "The Informativeness/Complexity Trade-Off in the Domain of Boolean Connectives". In: *Linguistic Inquiry*. DOI: 10.1162/ling_a_00461.
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.