Explaining gaps in the logical lexicon of natural languages

A decision-theoretic perspective on the square of Aristotle

Émile Enguehard[a] Benjamin Spector[a]

a. Institut Jean Nicod, Département d'Études Cognitives, École Normale Supérieure, PSL Research University, EHESS, CNRS, France

Keywords: language; decision-theoretic pragmatics; lexicon

Abstract

Across languages, certain logically natural concepts are not lexicalized, even though they can be expressed by complex expressions. This is for instance the case for the quantifier *not all*. In this paper, we propose an explanation for this fact based on the following idea: the logical lexicon of languages is partly shaped by a tradeoff between *informativity* and *cost*, and the inventory of logical expressions tends to maximize average informativity and minimize average cost. The account we propose is based on a decision-theoretic model of how speakers choose their messages in various situations (a modified version of the *Rational Speech Act* model).

1 Introduction

The Aristotelian square of opposition consists of the following four categories of logical statements:

1. *Universal (A):* All As are Bs.
2. *Negative universal (E):* No As are Bs.
3. *Existential (I):* Some As are Bs.
4. *Negative existential (O):* Some As are not Bs / Not all As are Bs.

An observation due to Horn (1973) is that English can express A, E and I more concisely than it can express O. Specifically, English can use a quantifying determiner to express A, E and I statements, but there is no corresponding word for O statements; instead one must use an additional negation. This can be seen in (1).

(1) a. *All* As are s. / *Every* A is a B.
b. *No* As are Bs.
c. *Some* As are Bs.
d. **Nall* As are Bs. / **Nevery* A is a B.

The general conclusion is that the abstract logical operators corresponding to A, E and I are (usually) *lexicalized* while O never is. This observation can be generalized to other languages as well as to temporal quantifiers and modal operators. For instance, 'not always' is less likely to be lexicalized than 'sometimes', 'never' and 'always' across languages, and 'unnecessary', which is lexicalized in English, is across languages less often lexicalized than 'possible', 'necessary' and 'impossible'.

At first sight, the observation is somewhat surprising, since the four operators all satisfy familiar constraints on possible quantifiers: not only conservativity, but also, in this case, monotonicity.[1] Furthermore, while O is the negation of A, I is the negation of E. And while I is the dual of A, O is the dual of E.

Horn (1973) provides an explanation why only three operators are needed: because they are logically related to one another (A entails I, E entails O, A and O and I and E are one another's negation), the four operators only allow us to distinguish between three basic situations or categories of worlds:

(2) $\forall$: Worlds where all As are Bs (A and I are true, E and O are false);
$\exists\neg\forall$: Worlds where some but not all As are Bs (I and O are true, E and A are false);
$\neg\exists$: Worlds where no As are Bs (E and O are true, A and I are false).

Furthermore, the mechanism of *scalar implicatures* lets the speaker indicate in which of these situations we are with just three operators. A and E statements can be used to indicate that we are in $\forall$ and $\neg\exists$ respectively. Then, if the speaker utters I, the listener may reason that A is false, as otherwise the speaker could have been more informative by saying A. For instance, an I statement like (3) is understood in most contexts to imply that the stronger A statement in (4) is false, or, equivalently, that we are in $\exists\neg\forall$; this is a well-known instance of the scalar implicature of an I statement. Thus, we just three operators, we are able to refer to the 3-way partition of worlds outlined in (2) with maximal precision; adding O to the mix would not allow us to be any more informative.

(3) Some of my colleagues are nice people.

(4) All of my colleagues are nice people.

The question, then, is why, in a given family of logical operators, it is always the set $\{A, E, I\}$ that is lexicalized, and not the other 3-element set $\{A, E, O\}$, with which a symmetric argument could be made. Indeed, O statements also trigger an inference to the effect that the corresponding E statement is false.

Horn (1973) proposes that E and O, being negative (downward-entailing), are marked in some sense and therefore dispreferred. Katzir and Singh (2013) generalize Horn's idea by proposing that at a certain level, logical operations are expressed in terms of certain primitives; this has the consequence that E and O have more complex representations. This is in line with findings suggesting that monotone-decreasing operators are harder to process than monotone-increasing ones (Geurts and Der Slik 2005).

Both Horn (1973) and Katzir and Singh (2013) essentially attempt to break the symmetry between I and O by assuming that O is inherently marked in some sense, making a lexicon that includes O rather than I dispreferred.

This note tries to approach the problem from a very different angle: under certain plausible assumptions, lexicalizing $\{A, E, I\}$ is optimal compared to $\{A, E, O\}$, in that it maximizes the *expected utility* that speakers can receive from using the language, where the utility of a single message in a single occasion of use depends on a trade-off between how informative the message is in this situation and how costly it is. We will make use of the same numerical notion of utility that is used in the Rational Speech Act model of pragmatics (Goodman and Stuhlmüller 2013, and Bergen, Levy, and Goodman 2016, where a cost-term is introduced in the utility function of messages).

On this approach the meaning of the operators, together with independently motivated assumptions about the meanings of lexical predicates and general principles of language use,

[1] Barwise and Cooper (1981) suggested that conservativity and monotonicity are semantic universals. See also Chemla, Buccola, and Dautriche (2018) for a recent discussion of a number of semantic properties, including monotonicity, which make some concepts more likely than others to be lexicalized across languages.

will let us derive an asymmetry.

We will proceed as follows: in Section 2, we will describe the general framework of utility-based models of pragmatics. In Section 3, we will demonstrate the basic integration of the square of Aristotle into such a model. In Section 4, we will introduce the notion of *expected utility*, which we need to consider global properties of the lexicon, and use it to derive a property of the denotation of predicates in natural languages: they denote, most of the time, a *minority* of the class of objects under consideration in a given sentence. Finally, in Section 5, we will provide a model of the expected utility of a language as a whole, and derive from the property of Section 4 that $\{A, E, I\}$, compared with $\{A, E, O\}$, is the lexicon that maximizes expected utility in our model.

2 Background: utility-based models of pragmatics

In so-called game-theoretic models of natural language pragmatics, speakers and listeners are seen as players in a collaborative signalling game; they aim to maximize a certain utility function, by following *strategies* that depend on their internal model of the other player. Such models include the *Rational Speech Act* model (RSA, Goodman and Stuhlmüller 2013) as well as the Iterated Best Response (IBR) and Iterated Quantal Response (IQR) models of Franke and Jäger (2014), among others. In what follows, we are going to adopt the formalism of RSA, though this will not be crucial, as the concepts we will use are common to all these models.[2]

Concretely, we assume that there is a certain set of possible worlds Ω, and a set of messages $\mathcal{M}$. The speaker knows exactly what the world is, while the listener's prior beliefs are represented by a probability distribution over Ω. The listener's prior beliefs are known to the speaker and more generally part of the Common Ground. Upon hearing a message, the listener updates their belief distribution. Thus, a model of the listener's behaviour is a function $L(w|m)$ giving the probability the listener assigns to world w after having heard message m.

Then, the utility of a message m for a speaker who knows the world is w and entertains a model L of the listener is given by $U_S(m|w; L)$:

(5) $U_S(m|w; L) = \log L(w|m) - \mathrm{cost}(m)$

The first term is the *informativity* term and corresponds to the fact that the speaker wants to maximize the listener's belief in w; the second term is the *cost* term and corresponds to the speaker's reluctance to utter complex messages. We assume that $\mathrm{cost}(m) \geq 0$ for all m.

Note that if the listener does not take message m to be compatible with w, that is, if $L(w|m) = 0$, then $U_S(m|w; L) = -\infty$. Thus a message which is *false* in the actual world (according to the listener's interpretation of the message) has minimal utility, and speakers will never use it.

We depart from Goodman and Stuhlmüller (2013) in assuming that speakers are maximally rational and always select the message that maximizes utility; in this respect our model is similar to IBR. This is for the sake of simplicity.[3] Also for the sake of simplicity, we ignore ties, i.e. cases where two or more messages maximize utility.[4] Thus, a model of the speaker is a function

[2] One feature specific to RSA that we rely on is its utility function defined purely in terms of listener belief. In IBR and IQR, the utility is taken to be a property of listener *actions*, which depend on listener beliefs in some way or other. From a purely formal point of view, this makes RSA a restricted case of IQR, which is obtained by identifying beliefs and actions. From a conceptual point of view, this means IQR and IBR's predictions are dependent on a notion of discourse purpose. Our argument does not involve any principled notion of discourse purpose other than belief update.

[3] This assumption makes calculations much simpler and lets us derive analytical results. However, there is no reason to think that our results would not hold in a model where the speaker isn't maximally rational.

[4] We ignore degenerate prior distributions where one world gets probability 1 and all others 0, in which case all messages compatible with the true world are ties. In the model discussed in section 3, there will be a tie between I ('some') and O ('not all') if the speaker believes both messages are true and the listener assigns to both the

S from worlds to messages; in each world, the speaker selects the message that maximizes their utility, given their model of the listener, as described in (6).[5]

(6) $S(w; L) = \arg\max_m U_S(m|w; L)$

One particular listener behavior is that of a *literal listener*. The literal listener L_0 has a prior distribution P_0 over worlds. They also have a notion of the semantics of each message: to each message m, they assign a set of worlds $[\![m]\!]$ where the message is true. Upon hearing m, the listener conditionalizes their belief distribution on m being true:

(7) $$L_0(w|m) = P_0(w|[\![m]\!]) = \begin{cases} 0 & \text{if } w \notin [\![m]\!], \\ \frac{P_0(w)}{P_0([\![m]\!])} & \text{if } w \in [\![m]\!]. \end{cases}$$

We can then define the first pragmatic speaker S_1 as the one whose behaviour is based on modeling the listener as L_0, and the first pragmatic listener L_1 as the one who models the speaker as S_1 and uses Bayes' rule to interpret their messages. Higher-order speakers S_2, S_3 etc. and listeners L_2, L_3 etc. may be defined similarly. We write $S_1(w)$ to denote the message that speaker S_1 chooses in world w; thus, $S_1(w) = S(w; L_0)$.

Goodman and Stuhlmüller (2013) show through numerical simulations that under reasonable assumptions on semantics and priors, for their RSA model, L_1 and higher-order listeners will draw conclusions similar to the scalar implicatures we observe for (3). We may then speculate that real-life speakers and listeners might behave like one of the higher-order speakers and listeners that we defined (though it is unclear which).

3 A utility-based model of the square of opposition

3.1 Deriving the agents' behaviour

In this section, we use the formalism of the previous section to build a model of speakers using the operators in the square of opposition to describe the world; similar models have been proposed by Franke (2009) and Goodman and Stuhlmüller (2013), among others, to demonstrate the capacity of such approaches to handle scalar implicatures.

Assume there is a relevant collection of $n > 1$ As, of which a certain subset are Bs. Recall that there are three possible situations or categories of worlds:

(8) $\forall$: All As are Bs.
$\exists\neg\forall$: Some but not all As are Bs.
$\neg\exists$: No As are Bs.

The speaker, having observed in which of these three situations, wants to tell the listener about it.

The three situations form a partition of the universe Ω. The messages the speaker may use are those of the square of opposition ($\mathcal{M} = \{A, E, I, O\}$), with their natural semantics:

same prior probability, distinct from 0 and 1 – a special case that we do not focus on. In the model discussed in section 5, the prior probability distribution over worlds is itself treated as a random variable. For certain values of this random variable, there will be a tie between I and O when they are both true. However, this only happens at a single point in the continuous probabilistic space, and the sort of expectation calculations that we carry out isn't affected by what happens at a single point, which is why we may simply ignore this possibility.

[5] It is important to keep in mind that while Goodman and Stuhlmüller (2013) also use S as a model of the speaker, their speakers are not maximally rational, which has the consequence that their S is a probability distribution over messages, given a world; thus it is a different mathematical object from our S. This change of notation lets us be more concise in some places.

(9)
$$\begin{aligned} [\![A]\!] &= \forall \\ [\![E]\!] &= \neg\exists \\ [\![I]\!] &= \exists\neg\forall \cup \forall \\ [\![O]\!] &= \exists\neg\forall \cup \neg\exists \end{aligned}$$

For the time being, we assume that all messages have the same cost. This means that we can ignore the cost term: since the cost term adds a constant to the utility of all messages, its actual value will not affect the comparison between messages.

Consider the first speaker S_1, whose behaviour is based on modelling the listener as the literal listener L_0. If the world w is in $\neg\exists$, i.e. if no As are Bs, then A and I are false and their utility is $-\infty$. We are left with E and O:

(10)
$$\begin{aligned} U_S(E|w; L_0) &= \log L_0(w|E) \\ &= \log \frac{P_0(w)}{P_0(\neg\exists)}, \\ U_S(O|w; L_0) &= \log L_0(w|O) \\ &= \log \frac{P_0(w)}{P_0(\neg\exists) + P_0(\exists\neg\forall)}. \end{aligned}$$

It is easy to verify that as long as:

(a) $P_0(w) > 0$ (we are not in a world that is incompatible with the common ground),

(b) $P_0(\exists\neg\forall) > 0$ (the common ground doesn't rule out the "some but not all" situation),

then E is a better message than O. Thus $S_1(w) = E$ whenever $w \in \neg\exists$.

Let's move on to the case where the world is in $\forall$, that is, all As are Bs. Messages E and O are false, and therefore will not be used. Under the same conditions as before, it is easy to check that A is always the better message. Thus $S_1(w) = A$ if $w \in \forall$. There is of course nothing surprising here: the model predicts that if all messages are equally costly, and one of the messages singles out the situation that the speaker wants to communicate, this message will be chosen.

The final and most interesting case is that of a world w in $\exists\neg\forall$: some, but not all As are Bs. This time, A and E are false and may not be used. I and O are both true in w but neither of them singles out w. Then:

(11)
$$\begin{aligned} U_S(I|w; L_0) &= \log \frac{P_0(w)}{P_0(\exists\neg\forall) + P_0(\forall)}, \\ U_S(O|w; L_0) &= \log \frac{P_0(w)}{P_0(\exists\neg\forall) + P_0(\neg\exists)}. \end{aligned}$$

Under the condition that $P_0(w) > 0$, the following holds:

(12) I is better than O as a message *if and only if* $P_0(\forall) < P_0(\neg\exists)$.

Equivalently:

(13) $S_1(\exists\neg\forall) = I$ *if and only if* $P_0(\forall) < P_0(\neg\exists)$.

We call C_0 the condition '$P_0(\forall) < P_0(\neg\exists)$'.

Now, this result captures the following intuition: if the actual world is in $\exists\neg\forall$, and one has a choice between saying I (i.e. 'some') and O (i.e. 'not all'), one will use the message whose truth is the most surprising given the listener's priors, i.e. the most informative in a probabilistic sense. Suppose for instance that condition C_0 is not satisfied, and instead we have $P_0(\neg\exists) < P_0(\forall)$. Then we also have $P_0(\neg\forall) < P_0(\exists)$, and therefore the proposition expressed by O (namely $\neg\forall$)

is more surprising, i.e. more informative, than that expressed by I ($\exists$). This is quite intuitive. Imagine that we are talking about an international scientific conference where it is expected that everybody will give her talk in English. If the speaker happens to know that, contrary to expectations, some talks will not be given in English but in French, using (14a) below seems much more appropriate than using (14b).

(14) a. Not every talk will be in English.
b. Some talks will be in English.

Correspondingly, (15b) seems much more appropriate than (15a) in the same situation (switching from English to French now makes condition C_0 true):

(15) a. Not every talk will be in French.
b. Some talks will be in French.

3.2 Interim discussion

The previous section has shown that, for a given prior distribution, only 3 messages will be used by S_1 (or any higher-order speaker); this is essentially the point already made by Horn (1973). Two of the messages will be A and E; the third one may be either I or O, depending only on a simple inequality. If it is more probable that no As be Bs than that all As be Bs, then speakers will say I; otherwise they will say O.

This first model already provides us with the beginning of an explanation for Horn's puzzle. A common intuition going back at least to Zipf (1929), who formulates it in the context of phonological change, is that the evolution of the lexicon is determined at least in part by a drive towards minimizing length and markedness: more frequent expressions will tend to be or become shorter and/or less marked than less frequent ones. Under the assumption that a single word like "some" is less marked than a compositional expression like "not all", Horn's observation that I is more often expressed by a single word than O would be expected if I is more frequent than O in language use. According to the model we've just developed, this ought to be the case if and only if C_0 holds in majority of discourse contexts. Thus, if we can find a reason why it would typically be the case that no As being Bs is more probable than all As being Bs (i.e. C_0 is typically true), then we could explain why natural languages prefer to lexicalize $\{A, E, I\}$ over $\{A, E, O\}$. This will allow speakers to use, most often, a short message, rather than a complex one (such as 'not all').[6] In section 4, we will argue that C_0 is in fact true most of the time, and will propose to derive this from considerations of utility.

There is, however, a small oddity in the reasoning we have just sketched. We said that "not all", being two words, is more marked than the single word "some". This suggests a notion of cost where "not all" ought to have greater cost than "some". However, in the model we used to derive Condition C_0, all messages have the same cost, and the choice of a message by the speaker only hinges on considerations of informativity. Thus, were we to formalize our reasoning in terms of global utility of a lexicon, we would need two notions of utility: speakers' behaviour

[6] The observation that frequent messages tend to be simple is also intuitively related to so-called manner implicatures. Manner implicatures (a.k.a. M-implicatures) is the phenomenon whereby a marked expression (e.g., 'cause to die') tends to refer to marked situations while unmarked ones (e.g., 'kill') tend to refer to prototypical situations. The question that M-implicatures raise is how expressions can have the same truth conditions and yet a different enriched meaning. Our problem is different: we want to explain an asymmetry between expressions with different truth conditions ("some" and "not all") but the same enriched meaning. Furthermore, in the case of manner implicatures, the marked expression specializes for a marked meaning, which corresponds to *unlikely* situations. In contrast, a key ingredient of our proposal will be that a meaning that corresponds to *likely* situations will be rarely used (because it is rarely informative enough), which favors lexica in which the meaning in question is not lexicalized.

As far as we can see, the accounts that have been proposed for manner implicatures in game-theoretic pragmatics (Van Rooij 2004; Benz and Van Rooij 2007; Franke and Jäger 2014; Bergen, Levy, and Goodman 2016, a.o.) cannot be straightforwardly extended to our case.

would be based on a notion of utility where cost plays no role, but as far as global utility is concerned, cost would matter.

In Section 5, we will resolve this discrepancy by providing a model in which the utility of a message in a given occasion includes a cost term that depends on its complexity, and where the overall utility of a lexicon is derived from the same utility function that determines usage. This model will also formalize the idea that prior distributions vary across occasions of uses and choices of predicates. We will formalize the intuition that if C_0 is usually true, then a lexicon based on $\{A, E, I\}$ gives rise to a higher expected utility than one based on $\{A, E, O\}$.

4 Expected utility and the optimality of predicates

The previous section assumed a fixed prior distribution P_0 on worlds. Realistically, in a given context, the prior probability of each of the three situations of interest ($\forall$, $\exists\neg\forall$ and $\neg\exists$) will depend upon what predicates A and B actually stand for and the specific features of the context. Thus, the optimality of some lexicon of operators over some other one will depend upon global properties of the lexicon of predicates.

In this section, we derive the following property of predicates: they are more likely to denote a minority of the kind of objects that they apply to than a majority. This captures the fact that we have, say, a word for dogs but not for non-dog animals, and that most lexical predicates, be they nouns or verbs, typically apply to a minority of the individuals that belong to a given super-category. This will follow from considerations of *expected utility*: defining words that way lets speakers achieve higher utility *on average*.[7]

The idea of using expected utility to break up symmetry in a model of this sort is already explored by Van Rooij (2004) among others to account for so-called manner implicatures; however the formal implementation of the idea and the question being addressed are very different from what we're doing here.

Qing and Franke (2014) also use expected utility to explain why gradable adjectives such as *tall*, whose denotation seems to involve an unspecified threshold, still receive a non-trivial interpretation: their idea is that the threshold is expected by speakers to be set in such a way as to maximize expected utility.[8]

In our case, we propose a model for a speaker who cares about classifying objects in a certain category A as being Bs or non-Bs. Let us take the word B to be defined only on As.[9] Having observed a new A, a speaker may want to tell others about it, and also whether it was a B. In this situation, the universe Ω is partitioned into two sets:

(16) NB: The A isn't a B.
B: The A is a B.

[7] Independently of considerations of informativity, there might be other reasons why we are more likely to lexicalize *dog* than *non-dog*: the *dog*-concept is arguably a more natural concept than the *non-dog* concept, for instance because it singles out a convex region of the conceptual space and is an intuitive sense more homogeneous (cf. Gärdenfors (2004)). What is really important to us is the generalization that words tend denote minorities, whatever the source of this tendency is. We note that in the few cases we could think of where a word divides a class into two categories which are intuitively as homogeneous and convex, the word tends to denote a minority of the class. This is true for gradable adjectives – 'tall boy' is understood to refer to boys who are significantly taller than average (cf. footnote 8) –, but also of pairs like *sick*/*healthy*, where in many languages *sick* is lexicalized but *healthy* isn't, or *round* vs. *non-round* (when talking about numbers).

[8] In fact, if one combines their idea with our definition of utility, it can be proven that a word like *tall* will optimally be true of a minority of a given comparison class.

[9] This is a simplification. In a more realistic model, As and Bs would be subclasses of say, Cs. Assuming that B denotes a minority of Cs, if subclasses of C that get their own word are reasonably widely distributed over subsets of C, then B ought to also denote a minority of most of them. Thus the conclusion that most of the time, B ought to denote a subset of A doesn't crucially depend on the assumption that B is defined on As.

We ignore here the possibility of compositionally complex messages such as “A but not B”.[10] Thus we assume there are two possible messages:

(17) A: “A!”
B: “B!”

Their semantics are the obvious ones:

(18)
$$[\![A]\!] = \Omega$$
$$[\![B]\!] = \textsc{b}$$

We assume that they have the same cost, which allows us to simply ignore the cost term again. It is straightforward to verify that S_1 will always say B in a world in B, and A in a world in NB. We can then compute the *expected utility* of the message used by S_1 when she encounters an A and says something about it. This quantity expresses how useful their utterance is on average. We assume again that the prior distribution is fixed and given by P_0. It represents both the actual probability that the A that S_1 encounters is a B, and the prior beliefs of the listener, who has not observed anything yet but has certain expectations before receiving information from S_1.

(19)
$$\begin{aligned}
\mathbb{E}_w[U_S(S(w;L_0)|w;L_0)] &= \sum_{w\in\textsc{b}} P_0(w)U_S(B|w;L_0) + \sum_{w\in\textsc{nb}} P_0(w)U_S(A|w;L_0)\\
&= \sum_{w\in\textsc{b}} P_0(w)\left(\log\frac{P_0(w)}{P_0(B)}\right) + \sum_{w\in\textsc{nb}} P_0(w)\left(\log\frac{P_0(w)}{P_0(A)}\right)\\
&= \overbrace{\sum_{w} P_0(w)\log P_0(w)}^{-H(P_0)} - \log P_0(B)\underbrace{\sum_{w\in\textsc{b}} P_0(w)}_{P_0(B)} - \log\underbrace{P_0(A)}_{1}\underbrace{\sum_{w\in\textsc{nb}} P_0(w)}_{1-P_0(B)}\\
&= -H(P_0) - P_0(B)\log P_0(B).
\end{aligned}$$

The first term does not depend on the lexicon: $H(P_0)$ depends solely on P_0, which is a parameter of the discourse context. However, the second term depends on what B means. In particular, imagine that speakers find themselves wanting to draw a new distinction within As, based on a specific binary feature. They could adopt a new word B^+, which refers to all As that have the feature one way, or B^-, referring to all As that have it the other way. This choice would have a consequence on the expected utility in (19), as the value of $P_0(B)$ wouldn't be the same: it could be either $P_0(B^+)$, or $P_0(B^-) = 1 - P_0(B^+)$. Then, we would expect speakers to actually adopt as B whichever term maximizes expected utility. We can verify that it is the rarer of the two: if $P_0(B^+) > 0.5$, it is optimal to choose B^- for B; if $P_0(B^+) < 0.5$, it is optimal to choose B^+ for B. Thus, in the end, it is always optimal to have $P_0(B) < 0.5$. We therefore expect that a concept B is more likely to be lexicalized than its negation $\neg B$ if, for most natural categories A within which distinguishing Bs from non-Bs would be relevant, a random A-individual is typically less likely to be a B than a non-B.

Before we move on to our second model, let us briefly return to Condition C_0. Recall that in Section 3.2, we provided a sketch of an explanation to Horn's puzzle, as long as we were able to justify that C_0 is true most of the time. We will now briefly see why that follows from our last result. Take a random individual x and two predicates A and B. We assume, following the previous discussion, that the priors are most often such that the probability p_0 that x is a B, given that that it is an A (i.e. $P_0(x \text{ is a B}|x \text{ is an A})$), is higher than the probability that x is not a B, given that it is an A. Equivalently, our assumption is that p_0 is lower than 0.5. If we

[10] A more complete model could integrate complex messages, but it would make our calculations much more complex. A simple approach could be to include them, but assign to them a prohibitive cost.

also assume that propositions of the form "*x is a B*", where x ranges over As, all have the same probability and are probalistically independent, we have, where n is the number of As:

(20)
$$P_0(\forall) = p_0^n$$
$$P_0(\neg\exists) = (1 - p_0)^n$$

It is easy to see that condition C_0 ($P_0(\forall) < P_0(\neg\exists)$) is satisfied if and only if $p_0 < 0.5$, which we have argued is expected to be most often the case. Then, as promised, it is also expected that C_0 holds most of the time. We will discuss the assumption of independence in greater detail in the context of our final model: in short, we will see that it is not principled, but also probably not crucial.

5 A model of the expected utility of a lexicon

In order to finally provide a satisfying explanation to Horn's puzzle, our goal in this section is to calculate the expected utility of the $\{A, E, I\}$ lexicon, as well as that of $\{A, E, O\}$. To do so, we need to account for variation in the As and Bs that speakers will talk about, and the diversity of the contexts in which they find themselves. This will be accomplished with two new elements, compared to our model of Section 3.

Our first new assumption has to do with the structure of the prior distribution P_0. We assume that every A has a certain probability p_0 of being a B, constant across As; furthermore, the propositions "this particular A is a B" form a set of independent events, in the sense of probability theory. As before, these simplifying assumptions let us calculate the relevant values of P_0 for each of the three situations at stake:

(21)
$$P_0(\forall) = p_0^n$$
$$P_0(\neg\exists) = (1 - p_0)^n$$
$$P_0(\exists\neg\forall) = 1 - p_0^n - (1 - p_0)^n$$

The simplifying assumptions we just made make P_0 dependent on a one-dimensional numerical parameter (p_0), which allows us to derive analytic results – more specifically, it allows us to work with a simple condition on p_0, stated below in (23). Nevertheless, these assumptions are not realistic. There is no reason to assume that every individual in a given domain has the same probability of having a certain property. And the independence assumption certainly does not generally hold. For instance, learning that Peter attended a concert could easily change the probability I would assign to Mary attending the concert, in different ways depending on the context, e.g. I might know that Mary and Peter are friends and that Peter is more likely to attend if Mary does. All we can say here is that these assumptions do not seem crucial to our argument. First, the fact that across contexts, these assumptions will be violated in all sorts of ways might let us hope that on average things work as if they held. Second, our model will depend solely on the value of $P_0(\neg\exists)$ and $P_0(\forall)$, so one could try to work directly with those, without relying on the parameter p_0.[11] The reason we don't do that is simply that it is more difficult to derive analytical results.

[11] Concretely, it seems plausible that the independence assumption will be violated in the following systematic way. Learning that an individual d in a set A has a property B will typically increase the probability that another individual d' in A has property B as well, because we think that A-individuals resemble each other; for the same reason learning that d doesn't have property B will decrease the probability that d' does. In other words, in many cases there will be a slight bias towards situations where all As are the same with respect to B (call this a *homogeneity bias*). This translates into both $P_0(\neg\exists)$ and $P_0(\forall)$ being higher than would result from the independence assumption. Now, in our model, everything depends upon which of $P_0(\neg\exists)$ or $P_0(\neg\forall)$ is the larger in a given context. Since, relative to the independence assumption, taking the homogeneity bias into account increases both probabilities, we can conjecture that which of them is the larger will typically depend on the value of p_0 as if the independence assumption held, and our conclusion will not be affected.

The second new element is that p_0, n and P_0 are not fixed, but are themselves random variables. This corresponds to the fact that across different contexts and different choices of predicates A and B, the prior conditional probability p_0 that a random individual d is a B, given that it is an A, is variable, as is the number of As. We take n and p_0 to be independent random variables: n ranges over the natural numbers greater than 2 and has probability g, and p_0 ranges over $[0, 1]$ and has density f.

Instead of assuming that the lexicon and the set of messages are the same thing, we will make the more realistic assumption that all four operators are possible messages, but that the one that isn't lexicalized has greater cost, because it takes a more complex expression to express it.[12] Thus we take the set of messages to be one of the following two possibilities:

(22) a. $\mathcal{M}_I = \{A, E, I, O^+\}$,
b. $\mathcal{M}_O = \{A, E, O, I^+\}$.

All messages have cost c, except those marked as $\Box^+$ which have cost $c^+ > c$. Thus, $\mathcal{M}_I$ corresponds to the lexicon where $\{A, E, I\}$ are lexicalized and it takes a compositional expression to say O (i.e. our lexicon[13]), while $\mathcal{M}_O$ is the lexicon where $\{A, E, O\}$ are lexicalized and it takes a compositional expression to say I (i.e. the one we want to rule out).

Finally, we assume that the inequality in (23) is true, which is a way of formalizing the property that we derived in Section 4: predicates are likely to be lexicalized in such a way as to make p_0 lower than 0.5. Intuitively, (23) expresses that the distribution of p_0 is biased towards the lower half of $[0, 1]$.[14]

(23) $\forall p_0 < 0.5, f(p_0) > f(1 - p_0)$.

We now have a model where we can compute the expected utility of a given language. Each value of p_0 and n determines a prior distribution P_0 over situations (world states), where situations are classified by which of the four Aristotelian statements they make true (we thus distinguish only three situations depending on whether no, some but not all, or all As are Bs). We assume that this prior distribution characterizes both the Common Ground belief and the probability, for each situation, that it is the actual situation. Given the prior distribution P_0, we know, for each situation s, how likely it is that s is the actual situation (this depends on P_0), which message the speaker uses in that situation, and the utility that is achieved. Averaging over all situations (where the weight of each situation depends on P_0), we can then compute the expected utility of the language for a given P_0. But P_0 is itself a random variable (defined in terms of the random variable p_0), and so we can now compute the expected utility of a language across all possible choices for P_0. This gives us the expected utility of the language as a whole.

As we'll formally prove in the appendix, given the condition in (23), the expected utility of $\mathcal{M}_I$ is greater than that of $\mathcal{M}_O$, the desired result. Before providing the precise mathematical result, let us explain at an intuitive level why it holds. It is based on the comparison of the utility achieved by each language in two types of situations where both I and O are true. Figure 1 provides an example of what the density of p_0 might look like, which helps visualize the comparisons.

1. **Uninteresting situations**.
   In cases where all As are Bs, speakers of both languages will systematically say A for the

[12] If we took only the lexicalized operators to be possible messages, our conclusion that $\{A, E, I\}$ is optimal would also follow; in fact calculations would be simpler.

[13] For our purposes, it doesn't matter whether O^+ corresponds to "some not", "not all" or any other way of saying O, and more generally what actual utterances messages correspond to: all that counts is that A, E, I are simpler.

[14] (23) is made true by a Beta law with parameters α and β where $\alpha < \beta$, which is arguably the most natural choice for a bell-shaped distribution over $[0, 1]$ whose mode is below 0.5. It is also made true by the distribution whose density is proportional to (an increasing function of) the expected utility of p_0, as given by (19).

reasons outlined in section 3.1. Similarly when no As are Bs, they will always say E. Thus these situations play no role in the comparison of the two lexica. We are only interested in what happens when some but not all As are Bs.

2. **Situations where the more costly message is used.**
In both languages, the costly message is used in "some but not all" situations only if it is highly informative compared to the less costly one, so that the disadvantage it has in terms of cost is overridden. Now, in $\mathcal{M}_O$, where the costly message is I, this will happen when the prior probability of $\neg\exists$ is sufficiently high (so that the costly message, which means $\exists$, will be highly informative), that is, when p_0 is sufficiently low; this corresponds to the left-hand region on Figure 1. Meanwhile, in $\mathcal{M}_I$, this will happen when the prior probability of $\forall$ is sufficiently high (so that the costly message, which means $\neg\forall$, is highly informative): this corresponds to the right-hand region on Figure 1. Given our assumption that, p_0 takes smaller values more often, we will be more often in this situation in $\mathcal{M}_O$ than in $\mathcal{M}_I$ — on Figure 1, this is seen by the fact that the left-hand region has a larger area. So the more costly message will be used more often in $\mathcal{M}_O$ than in $\mathcal{M}_I$, which creates a disadvantage for $\mathcal{M}_O$ on the cost side.

3. **Situations where the less costly message is used.**
When the prior P_0 is not sufficiently biased so as to make the costly message optimal, speakers always use the simpler message in "some but not all" cases (i.e. I in $\mathcal{M}_I$ and O in $\mathcal{M}_O$). This corresponds to the middle area in Figure 1. In this case, the comparison between the two languages hinges on how informative the simpler message is relative to P_0: when P_0 is such that C_0 holds, the most informative message is I, and speakers of $\mathcal{M}_O$, who say O, incur a loss of utility. When the priors are such that C_0 does not hold, the most informative message is O, and speakers of $\mathcal{M}_I$, who say I, incur a symmetric loss of utility. Now, because of our assumption that C_0 holds most of the time, which remains true when one restricts oneself to less biased priors, the former situation will be more frequent than the latter; on Figure 1, this is seen by the fact that the left-hand half of the middle area is larger than the right-hand half. Essentially, $\mathcal{M}_I$ will sometimes lead to diminished informativeness, but this will occur less often than equivalent losses under $\mathcal{M}_O$. Here again, there is an advantage for $\mathcal{M}_I$, this time on the informativity side.

The actual calculation of expected utility for each set of messages is given in Appendix A. The difference between the expected utility for $\mathcal{M}_I$ and that for $\mathcal{M}_0$ is given by ΔU below:

(24)
$$\Delta U = (c^+ - c)\Big(\mathbb{P}(\Phi(P_0) > c^+ - c) - \mathbb{P}(\Phi(P_0) < c - c^+)\Big) \\ + \mathbb{E}_{P_0}\Big[\Phi(P_0) \mid c - c^+ < \Phi(P_0) < c^+ - c\Big],$$

where Φ is a certain function defined in the Appendix. Without going into the details, the first term corresponds to the extra cost that speakers incur when they use a complex expression in each lexicon (second type of situations above). The second term corresponds to the loss of informativity that speakers incur when they forsake a little bit of informativity to save cost (third type of situations above). We show in the Appendix that it follows from Condition (23) that both terms are positive. Thus, as long as (23) holds, $\Delta U > 0$, and the lexicon corresponding to $\mathcal{M}_I$ is optimal.[15]

[15] Our calculations look at the optimal utility of the speaker S_1. Instead, one could have looked at the utility of S_2. Because S_2 models the listener as L_1, and L_1 derives quantity implicatures, L_1 treats I and O as meaning "some but not all" whenever they are used. As a result, whenever S_2 chooses I or O, S_2 is always maximally informative. The consequence for ΔU is that the second term disappears. Because the first term is already positive, ΔU is still positive, so our argument would still go through. However, the argument of Section 4 wouldn't have worked as is with S_2: lexicalizing B^+ or B^- would provide the same S_2-utility.

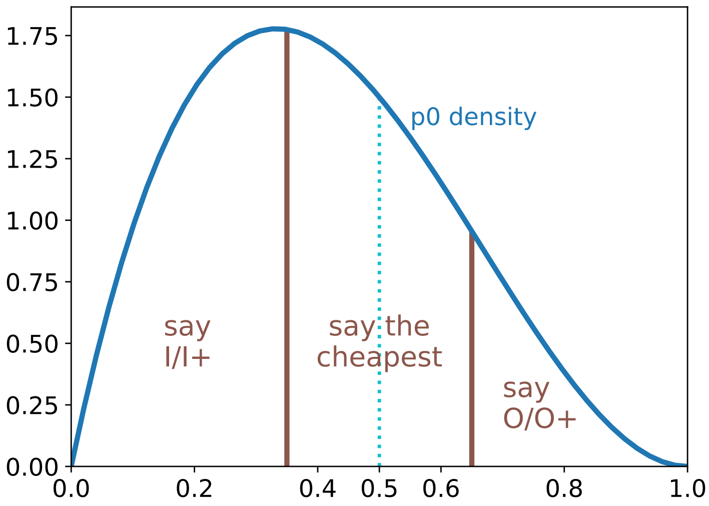

Figure 1: An example of a density f biased towards smaller values (in this case $\mathrm{Beta}(2,3)$). The x-dimension represents possible values of p_0; it is divided into three intervals. In the first interval, speakers of both languages say I or I^+; in the second one, speakers of both languages say the cheaper message; in the third one, they say O or O^+. The probability of each of these three cases is given by the area under the curve. The dotted line marks the limit between the domain where C_0 holds and I is most informative in $\exists\neg\forall$ situations, and the domain where the more informative message is O.

6 Discussion

Let us take stock. Having to draw distinctions within A, considerations of utility will lead to words denoting a minority (<50%) of As to be lexicalized preferentially. Under a specific formalization of this fact (Condition (23)), it can be proved that the expected utility provided by a lexicon where $\{A, E, I\}$ are cheap and O is expensive is higher than the expected utility provided by a lexicon where $\{A, E, O\}$ are cheap and I is expensive. This might explain why our actual lexicon is closer to the first one.

In order to derive this result, we made a number of simplifying assumptions: full rationality of the speakers, simple structure of P_0, etc. While none of them seem crucial, they make calculations easier and let us derive analytical results. Independently of these simplifying assumptions and our specific modeling choices, our results in Sections 3 and 4 suggest the following story. English has a word for bakers but no word for people who don't sell bread, it has a word for dogs but no word for non-dog animals, and so on. Section 4 offers an account for why this is expected. It then intuitively follows, even without our specific assumptions, that Condition C_0 will be true most of the time. As shown in Section 3, this will lead to I being used more than O, and we will see it being lexicalized first — but the causal connection between 'being used more' and 'being lexicalized first' is not itself part of the model discussed in these sections. In Section 5, we offer a more explicit approach where we compare the overall expected utility of different lexica, but to do so we needed to make some more specific assumptions about the structure of priors.[16]

Our approach is limited in scope in that we only compare $\{A, E, I\}$ to $\{A, E, O\}$, but not to lexica with a different numbers of lexicalized corners of Artistotle's square. In particular, we do not really explain why lexicalized O is so rare. In principle, one might think that $\{A, E, I, O\}$ would be the ideal lexicon: speakers do feel the need to make O statements from time to time.[17]

[16] Note also that we assumed that the speaker is always maximally informed.

[17] In the model discussed in section 3, they do so precisely when C_0 is false, i.e. when A is quite probable or E is quite improbable *a priori*. In the model discussed in section 5, the difference between the two quantities compared in condition C_0 has to be larger than some positive number, but the end-result is qualitatively similar.

We need to assume that some independent pressure to keep the lexicon minimal prevents the lexicalization of all four items. Ideally, this pressure would be part of our model.[18]

Similarly, one may wonder what is the expected utility of lexicalizing fewer corners. Thus, we would hope that lexicalizing $\{A, I\}$ is optimal within the two-element lexica, since this seems to be what is most common in natural languages (Katzir and Singh 2013). Taking compositionality into account (one can construct the missing messages by adding a negation), we would then compare the expected utilities of $\{A, E^+, I, O^+\}$, $\{A^+, E, I^+, O\}$ and $\{A, E, I^+, O^+\}$. It turns out that this time, there are terms of differing signs in the differences, and our current assumptions do not let us conclude as to the overall sign. Our approach therefore does not let us decide between these lexica, at least in its present form.

Finally, as we noted, the observation that I is more likely to be lexicalized than O holds not only in the domain of quantifiers over individuals, but also in the temporal and modal domains, as discussed by Horn (1973). Our approach would not have much difficulty to generalize to such cases, on the plausible assumption that, on average, the argument of such modal and temporal operators denote propositions which have a lower prior probability than their negation. Any more sophisticated method accounting for the size of the lexicon would have to handle the fact that in certain domains, all four operators appear to be lexicalized.[19]

7 Conclusion

In this paper, we have seen how a decision-theoretic model of pragmatics lets us break the symmetry between the four corners of the logical square of opposition: given a prior distribution on worlds, there will always been one of the I or O corners which is better than the other in terms of *utility*, where utility is defined as in the RSA model. Which is chosen depends solely on a simple condition on prior distributions. We then argued that it follows that lexicalizing I leads to higher utility on average, which explains why I is indeed more prominent in natural language; our argument builds upon the assumption that most predicates usually denote a minority of the domains over which they are used, a fact that itself also follows from considerations of expected utilty. We developed a model that relied on a number of simplifying assumptions. These simplifying assumptions allowed us to prove our results analytically instead of simulating the results of a more complex model. It remains to be seen whether the argument we gave can be maintained in a more complex and realistic model. However conclusive our argument may appear in this particular case, we hope that this paper provides an example of how explicit decision-theoretic models of pragmatics can in principle account, through the notion of expected utility, for certain universal tendencies in the lexicon of natural languages.

8 Acknowledgements

The research leading to these results has received funding from the Agence Nationale de la Recherche (ANR-10-IDEX-0001-02). The authors thank Emmanuel Chemla for useful discussions and comments.

[18] One could argue that when one takes into account their scalar implicatures, I and O statements are truth-conditionally equivalent, in that they denote the $\exists\neg\forall$ situation. Therefore, there is no sense in lexicalizing operators for both; in fact, Aristotle's identification of four basic statements might be less relevant to natural language than the natural partition of possible worlds into just three sets. This is essentially the line of argument of Horn (1973, pp. 251-260). However, as we pointed out, O statements are widely attested and are not used in the same contexts as I statements (cf. our discussion of examples (14) and (15) in section 3.1); for this reason, we are reluctant to just "shave off [the O category] with Occam's razor" (Horn 1973, p. 259).

[19] An English example is modal adjectives: *possible* (I), *impossible* (E), *necessary* (A) and *unnecessary* (O).

References

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

Benz, Anton and Robert Van Rooij (2007). "Optimal assertions, and what they implicate. A uniform game theoretic approach". In: *Topoi* 26.1, pp. 63–78.

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

Chemla, Emmanuel, Brian Buccola, and Isabelle Dautriche (2018). "Connecting content and logical words". In: *Journal of Semantics.*

Franke, Michael (2009). "Signal to act: Game theory in pragmatics". Ph.D. dissertation, ILLC, Universiteit van Amsterdam.

Franke, Michael and Gerhard Jäger (2014). "Pragmatic Back-and-Forth Reasoning". In: *Pragmatics, Semantics and the Case of Scalar Implicatures.* Ed. by Salvatore Pistoia Reda. London: Palgrave Macmillan UK, pp. 170–200.

Gärdenfors, Peter (2004). *Conceptual spaces: The geometry of thought.* MIT press.

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

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.

Horn, Laurence Robert (1973). "On the semantic properties of logical operators in English". Ph.D. dissertation, UCLA.

Katzir, Roni and Raj Singh (2013). "Constraints on the lexicalization of logical operators". In: *Linguistics and Philosophy* 36.1, pp. 1–29.

Qing, Ciyang and Michael Franke (2014). "Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model". In: *Semantics and linguistic theory.* Vol. 24, pp. 23–41.

Van Rooij, Robert (2004). "Signalling games select Horn strategies". In: *Linguistics and philosophy* 27.4, pp. 493–527.

Zipf, George Kingsley (1929). "Relative Frequency as a Determinant of Phonetic Change". In: *Harvard Studies in Classical Philology* 40, pp. 1–95.

A Calculation of the expected utility

This Appendix contains the calculations of the expected utility for the models in Section 5, and the derivation of ΔU.

First, we define the following function Φ, which will prove convenient:

(25) $\Phi(P_0) = \log \frac{P_0(\exists\neg\forall)+P_0(\neg\exists)}{P_0(\exists\neg\forall)+P_0(\forall)} = \log \frac{1-p_0^n}{1-(1-p_0)^n}.$

$\Phi(P_0)$ can be thought of as the difference in informativity between saying I and saying O, when they are both true. Viewed as a function of p_0 (for a given n), it has the following properties:

(26) a. $\Phi(p_0) > 0$ if and only if $p_0 < 0.5$,
b. $\Phi(p_0) = -\Phi(1-p_0)$ for all p_0.
c. Φ is monotone decreasing over $[0, 1]$.

Under both sets of messages, $\mathcal{M}_I$ and $\mathcal{M}_O$, speakers will always say A or E when they are true, for the same reasons as in Section 3. When neither is true, we are in the $\exists\neg\forall$ situation. Then, they will say one of $I^{(+)}$ or $O^{(+)}$, depending on the following conditions:

(27) a. If the lexicon is $\mathcal{M}_I$, speakers say I if and only if $\Phi(P_0) > c - c^+$.

b. If the lexicon is $\mathcal{M}_O$, speakers say I^+ if and only if $\Phi(P_0) > c^+ - c$.

Because $\Phi(P_0)$ increases with $P_0(\neg\exists)$ and decreases with $P_0(\forall)$, these conditions are qualitatively similar to C_0: they are true when $\neg\exists$ is highly probable and false when $\forall$ is highly probable. C_0 sets the threshold for $\Phi(P_0)$ at 0. Once differences in cost enter the pictures, under each lexicon, the threshold is either above or below 0. This corresponds to the fact that speakers are willing to lose a bit of informativity in order to save on cost.

We can now compute the expected utilities in each system:

(28)
$$\begin{aligned}\mathbb{E}_{P_0}\left[\mathbb{E}_w\left[U_S(\dots)\right]\right] &= \mathbb{E}_{P_0}\left[\mathbb{E}_w\left[\log\frac{P_0(w)}{P_0(S_1(w))} - \mathrm{cost}(S_1(w))\right]\right]\\ &= \underbrace{\mathbb{E}_{P_0}\left[\mathbb{E}_w\left[\log P_0(w)\right]\right]}_{\text{call this } K} - \sum_{s\in\{\forall,\neg\exists,\exists\neg\forall\}}\mathbb{E}_{P_0}\left[P_0(s)(\log P_0(S_1(s)) + \mathrm{cost}(S_1(s)))\right]\end{aligned}$$

For $\mathcal{M}_I$:

(29)
$$\begin{aligned}\mathbb{E}_{P_0}\left[\mathbb{E}_w\left[U_S(\dots)\right]\right] = K &- \mathbb{E}_{P_0}\left[P_0(\forall)\log P_0(\forall) + c\right] - \mathbb{E}_{P_0}\left[P_0(\neg\exists)\log P_0(\neg\exists) + c\right]\\ &- \mathbb{E}_{P_0}\left[\mathbb{1}_{[\Phi(P_0)>c-c^+]}(\log(P_0(\forall) + P_0(\exists\neg\forall)) + c)\right]\\ &- \mathbb{E}_{P_0}\left[\mathbb{1}_{[\Phi(P_0)<c-c^+]}(\log(P_0(\neg\exists) + P_0(\exists\neg\forall)) + c^+)\right]\end{aligned}$$

For $\mathcal{M}_O$:

(30)
$$\begin{aligned}\mathbb{E}_{P_0}\left[\mathbb{E}_w\left[U_S(\dots)\right]\right] = K &- \mathbb{E}_{P_0}\left[P_0(\forall)\log P_0(\forall) + c\right] - \mathbb{E}_{P_0}\left[P_0(\neg\exists)\log P_0(\neg\exists) + c\right]\\ &- \mathbb{E}_{P_0}\left[\mathbb{1}_{[\Phi(P_0)>c^+-c]}(\log(P_0(\forall) + P_0(\exists\neg\forall)) + c^+)\right]\\ &- \mathbb{E}_{P_0}\left[\mathbb{1}_{[\Phi(P_0)<c^+-c]}(\log(P_0(\neg\exists) + P_0(\exists\neg\forall)) + c)\right]\end{aligned}$$

We can calculate the difference ΔU; most terms simplify out:

(31)
$$\begin{aligned}\Delta U = (c^+ - c)&\Big(\mathbb{P}(\Phi(P_0) > c^+ - c) - \mathbb{P}(\Phi(P_0) < c - c^+)\Big)\\ &+ \mathbb{E}_{P_0}\left[\Phi(P_0) \mid \quad |\Phi(P_0)| < c^+ - c\right]\end{aligned}$$

Here is an intuitive explanation of the terms:

a. $(c^+ - c)\mathbb{P}(\Phi(P_0) > c^+ - c)$ is the difference in cost from cases where speakers of both lexica say I/I^+: those who say I^+ incur higher cost.

b. $(c - c^+)\mathbb{P}(\Phi(P_0) < c - c^+)$ is the difference in cost from cases where speakers of both lexica say O/O^+: those who say O^+ incur higher cost.

c. $\mathbb{E}_{P_0}\left[\Phi(P_0) \mid \quad |\Phi(P_0)| < c^+ - c\right]$ is the average loss of informativity in cases where Φ is close to 0 ($|\Phi(P_0)| < c^+ - c$, the absolute difference in informativity between I and O is small) and speakers just say the cheapest thing, whether it is the most informative or not.

To conclude, we have to derive that $\Delta U > 0$ from Condition (23).

(a) $c^+ - c > 0$ by assumption.

(b) $$\begin{aligned}\mathbb{P}(\Phi(P_0) > c^+ - c) &= \sum_n g(n) \int_{\Phi(p_0)>c^+-c} f(p_0)\mathrm{d}p_0 \\ &= \sum_n g(n) \int_{\Phi(1-p_0)<c-c^+} f(p_0)\mathrm{d}p_0 \\ &> \sum_n g(n) \int_{\Phi(1-p_0)<c-c^+} f(1-p_0)\mathrm{d}p_0 \\ &> \sum_n g(n) \int_{\Phi(p_0')<c-c^+} f(p_0')\mathrm{d}p_0' \\ &> \mathbb{P}(\Phi(P_0) < c - c^+).\end{aligned}$$

(c) $$\begin{aligned}\mathbb{E}_{P_0}\left[\Phi(P_0) \mid |\Phi(P_0)| < c^+ - c\right] &= \sum_n g(n) \int_{|\Phi(p_0)|<c^+-c} f(p_0)\Phi(p_0)\mathrm{d}p_0 \\ &= \sum_n g(n) \left(\int_{c-c^+<\Phi(p_0)<0} f(p_0)\Phi(p_0)\mathrm{d}p_0 \right. \\ &\qquad \left. + \int_{0<\Phi(p_0)<c^+-c} f(p_0)\Phi(p_0)\mathrm{d}p_0\right) \\ &= \sum_n g(n) \left(\int_{0<\Phi(p_0)<c^+-c} f(p_0)\Phi(p_0)\mathrm{d}p_0 \right. \\ &\qquad \left. - \int_{c-c^+<\Phi(p_0)<0} f(p_0)\Phi(1-p_0)\mathrm{d}p_0\right) \\ &> \sum_n g(n) \left(\int_{0<\Phi(p_0)<c^+-c} f(p_0)\Phi(p_0)\mathrm{d}p_0 \right. \\ &\qquad \left. - \int_{c-c^+<\Phi(p_0)<0} f(1-p_0)\Phi(1-p_0)\mathrm{d}p_0\right) \\ &> 0.\end{aligned}$$

From these three inequalities, it follows that $\Delta U > 0$. Hence, $\mathcal{M}_I$ provides higher expected utility as long as Condition (23) holds (which is a weak formalization of the idea that most predicates are minorities) is respected.