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

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 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, this 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 and A are each other's negation, and one might think that this would make having both unnecessary, I and E are each other's negation. Similarly, while I is the dual of A, O is the dual of E. The difference in terms of lexicalization between I and O is therefore unexpected.

Horn (1973) suggested 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)
$w_{\forall}$: Worlds where all As are Bs (A and I are true, E and O are false);
$w_{\exists\neg\forall}$: Worlds where some but not all As are Bs (I/O are true, E/A false);
$w_{\neg\exists}$: 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 $w_{\forall}$ and $w_{\neg\exists}$ 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 $w_{\exists\neg\forall}$; 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.

Horn's idea, however, does not explain why, in a given family of logical operators, it is always the set $\{A, E, I\}$ which 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. Furthemore, as we will discuss, speakers sometimes do use O statements, and there are contexts where O statements are more felicitous than I statements (cf., e.g., the contrast in (12) below).

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

[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.

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. Katzir and Singh (2013) implement this idea by assuming a specific inventory of cognitive primitives such that monotone increasing operators have simpler representations. We should note however, that in the absence of such hypotheses about cognitive primitives, there is no reason to believe that I is intrinsically computationally less complex that O. Consider for instance a measure of complexity based on the semantic automata that can represent a given quantifier (see, e.g., Steinert-Threlkeld and Szymanik 2019; Katzir, Lan, and Peled 2020). An automaton for I takes a sequence of 0s and 1s and returns 'true' if the sequence contains at least one occurrence of 1. It is trivial to convert such an automaton into one with exactly the same structure which returns 'true' if the sequence contains at least one occurrence of 0, thus encoding the meaning of O. As to the observed cognitive cost of sentences containing negation and other monotone-decreasing operators, it should be noted that such sentences tend to be *syntactically* more complex than their 'positive' counterparts, so that we should not conclude that this cost is specifically tied to the logical meaning of the relevant operators. We discuss these points with more details in section 4.2, in connection with recent proposals that are to some extent related to ours.[2]

We approach the problem from a very different angle. First we will argue that there are principled reasons why, irrespective of which corners of the Aristotelian square are lexicalized, O-statements are expected to be less frequently used than I-statements. Given the well-established relation between frequency and lexicalization (the more frequently an expression is used, the more likely it is to be lexicalized), this might provide an explanation for why I tends to be lexicalized but O does not. Second, we will show that 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. The account we will propose is in line with the view that a number of features of natural languages can be understood as maximizing the overall utility of a language (cf. Gibson et al. 2019 and the references cited therein). We will make use of the same information-theoretic definition of utility that is used in the Rational Speech Act model of pragmatics (RSA; Goodman and Stuhlmüller 2013, and Bergen, Levy, and Goodman 2016, where a cost-term is introduced in the utility function of messages), but our account is otherwise not couched in a game-theoretic framework. For both lines of explanation, a crucial ingredient of our account is the observation (already made in Chater and Oaksford 1999) that, on average, an O-statement (e.g., 'Not all of the guests were drunk') is less informative than its corresponding I-statement (e.g. 'Some of the guests were drunk').

On our approach the asymmetry will be derived directly from the truth conditions of the operators, together with independently motivated assumptions about the meanings of lexical predicates and general principles of language use. In particular, cognitive or morphological primitives will play no role in our explanation.[3]

[2]See, among others, Penka (2011), Zeijlstra (2011), and Buccola and Spector (2016), for arguments that monotone-decreasing quantifiers such as *no* or *fewer than 10* are to be syntactically decomposed into a negation-like operator and an upward-entailing quantifier, based in part on the availability of so-called 'split-scope' readings.

[3]Several recent works (Denić, Steinert-Threlkeld, and Szymanik 2021; Uegaki 2020, a.o.) also explore the inventory of lexical quantifiers from an information-theoretic perspective, but their approach relies

We will proceed as follows: in Section 2, we will explain why I-statements are expected to be used more frequently that O statements. In section 3, we provide a model of the expected utility of a lexicon, in which the expected utility of a lexicon based on $\{A, E, I\}$ has a greater expected utility than one based on $\{A, E, O\}$.

2 Why I is expected to be more frequent than O

It would be no surprise to find, in a corpus study, that I-sentences of the form *Some of the NPs VP* (e.g., *Some of the guests were drunk*) have on average more occurrences than corresponding O-sentences of the form *Not All of the NPs VP* or *The NPs AUX not all VP* (e.g., *Not all of the guests were drunk*, *All the guests were not drunk*). Given that, in English and in other languages, O is not lexicalized, O-sentences are bound to be syntactically more complex than I-sentences, which would be enough to explain the observed difference in frequency of occurrences, on the plausible assumption that, everything else being equal, more complex constructions are less frequent than less complex ones. Such a corpus-based observation would not provide any convincing explanation for the fact that I is lexicalized and O is not, since the lexicalization facts might in fact explain the observed frequencies.

What we want to argue here is that even if the four corners of the Aristotelian square were lexicalized, O-sentences would still be less frequently used than I-sentences.

Imagine a speaker who knows that, say, some but not all of the guests at a certain party were drunk. The two 'universal' corners of the square cannot be used, as they yield false statements. If the speaker is to use one of the four quantifiers of the square, the choice to make is, then, between the following sentences:

(5) a. Some of the guests were drunk.
b. Not all of the guests were drunk.

Now, both sentences trigger a 'some but not all' implicature, so once this is taken into account they should be equally effective. If we only consider the literal meaning of both sentences, and use a notion of informativity based on entailment, neither sentence is more informative than the other, since there is no entailment relation in either direction. On such a view of informativity, Grice's maxim of Quantity cannot help us choose between the two sentences. However, suppose that we use instead a probabilistic notion of informativity, where, in a given context, a proposition ϕ is more informative than a proposition ψ if, before she hears the sentence, the hearer considers ϕ less likely than ψ. In that case, we have to ask which of the two propositions in (5) is the least likely to be true in the context of utterance, from the point of view of the listener. In many contexts, the probability that no guest is drunk (i.e. the probability of the negation of (5a)) is higher than the probability that all guests are drunk (i.e. the probability of the negation of (5b)). Equivalently, in such contexts, (5a) is less likely to be true than (5b), hence, given the perspective we adopt here, more informative. Assuming that speakers tend to prefer more informative sentences, (5a) will be used in a such a context.

Suppose now that, in most contexts, I-statements are more informative than O-statements in this sense. Then we would predict that, irrespective of lexicalization, I

on specific assumptions on lexical primitives or meaning spaces similar to those made by Katzir and Singh (2013). We will discuss the relevance of these analyses to Horn's puzzle and their similarities and differences with our proposal in Section 4.2.

would be more frequently used than O. Such a prediction might in turn help explain the lexicalization facts.

To substantiate an account of this type, we need to a) establish that indeed the choice between I and O is partly governed by probabilistic informativity, and b) that there are good reasons to think that across contexts I is indeed more informative (in this sense) than O.

Before turning to these issues, we will provide a model of message choice which captures the reasoning we have just sketched, inspired by the RSA model of pragmatics (Goodman and Stuhlmüller 2013).

2.1 A simple model of the pragmatics of the Aristotelian square

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. 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:

$$L_0(w|m) = P_0(w|[\![m]\!]) = \begin{cases} 0 & \text{if } w \notin [\![m]\!], \\ \frac{P_0(w)}{P_0([\![m]\!])} & \text{if } w \in [\![m]\!]. \end{cases} \tag{6}$$

We assume for the time being that the speaker chooses her message m in the following manner: if they are in world w, they pick the message that maximizes the probability that the listener will assign to w after processing it.[4] We write $S(w; L)$ to refer to the message chosen by a speaker who believes w.[5]

$$S(w; L) = \arg\max_m P_0(w|m) \tag{7}$$

Now, let us assume that worlds are individuated by whether all ($w_\forall$), some but not all ($w_{\exists\neg\forall}$), or no ($w_{\neg\exists}$) guest is drunk. Let us assume that there are four messages: A (All guests are drunk), E (No guest is drunk), I (Some guests are drunk), O (Not all guests are drunk). Now, if in fact no guest is drunk, it is obvious that the best message is E, since after processing it the listener assigns 1 to the world where no guest is drunk (and the three other messages fail to achieve the same effect). Similarly if in fact all guests are drunk, the best message is A. The interesting case is the some-but-not-all world (denoted

[4] Here we depart from the standard RSA model, where the speaker is not fully rational, and does not always pick the best message, but, rather, picks each message with a probability which is increasing with the informativity of the message. In this particular respect, the model presented in this section is similar to the earlier Optimal Answer model of Benz and Van Rooij (2007) and to Franke's (2011) Iterated Best Response Model, among others. We make this choice because we think that it will make our proofs clearer and more intuitive later on; it is not crucial, and would not affect our conclusions. See also Footnote 16.

[5] In principle, two messages could be exactly tied and be both optimal, so $S(w; L)$ is not a function. In our model, this will happen when the prior probability distribution P_0 of the listener is such that $P_0(s_\exists) = P_0(s_{\neg\forall})$ For simplicity, we ignore the possibility of such a tie, i.e. do as if such a prior distribution were not possible, which does not affect our conclusions in any way.

by $w_{\exists\neg\forall}$). In such a situation the two messages that could be used are I and O. Given (7), the speaker will choose I if and only if the following holds:

(8) $L(w_{\exists\neg\forall}|I) > L(w_{\exists\neg\forall}|O)$

Now:

(9) a. $L(w_{\exists\neg\forall}|I) = P_0(w_{\exists\neg\forall}|[\![I]\!]) = P_0(w_{\exists\neg\forall}|\{w_{\exists\neg\forall}, w_\forall\}) = \frac{P_0(w_{\exists\neg\forall})}{P_0(w_{\exists\neg\forall})+P_0(w_\forall)}$

b. $L(w_{\exists\neg\forall}|O) = P_0(w_{\exists\neg\forall}|[\![O]\!]) = P_0(w_{\exists\neg\forall}|\{w_{\exists\neg\forall}, w_{\neg\exists}\}) = \frac{P_0(w_{\exists\neg\forall})}{P_0(w_{\exists\neg\forall})+P_0(w_{\neg\exists})}$

It then follows straightforwardly that:

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

Equivalently (with $s_\exists = \{w_{\exists\neg\forall}, w_\forall\}$ and $s_{\neg\forall} = \{w_{\exists\neg\forall}, w_{\neg\exists}\}$)

(11) I is better than O as a message *if and only if* $P_0(s_{\neg\forall}) > P_0(s_\exists)$,
i.e. if and only if $P_0([\![O]\!]) > P_0([\![I]\!])$

This was the expected result: the speaker chooses the message which expresses the proposition that was the least likely to be true given the prior distribution, i.e, whose *surprisal value* is the highest.

Now, as noted by a reviewer, this prediction crucially relies on the view that the speaker, when choosing her message, measures its informativity (surprisal value) in terms of its literal interpretation. That is, we are discussing here the behavior of the fully rational version of the level-1 Speaker of the Rational Speech Act model, who assumes she is talking to a literal listener. A more sophisticated speaker might assume that she talks to a pragmatic listener. Because the pragmatic listener would interpret both 'some' and 'not all' as meaning 'some but not all', both are going to be equally informative, and our proposal in this paper could not work if we modeled the speaker in this way. We think we can argue for our choice on the following grounds. First, as discussed in the next section, it seems to be a fact that the choice between 'some' and 'not all' is partly governed by the relative informativity (measured in terms of surprisal) of their literal meaning (cf. our discussion of the examples in (12) and (13) below). Second, scalar implicatures are in any case not always derived by the listener (in the experimental literature on scalar implicatures, rates of scalar implicature derivation are never very close to 100%, and are typically lower than for prototypical entailments). If the speaker believes that there is a small chance that she is talking to a literal listener (or a listener who believes that the speaker is not knowledgeable about the alternative, or one who takes the relevant alternative - say 'all' in the case of 'some' - to be irrelevant), she will always be better off choosing the message whose literal meaning is the most informative.[6]

[6]We could of course consider a more sophisticated model where the speaker is uncertain as to whether she talks to the literal listener or to a pragmatic listener (or as to the value of other parameters, such as the underlying Question Under Discussion the listener is entertaining), but qualitatively the outcome would be the same. Note also that if we use the 'standard' RSA model where speakers are only approximately rational, the effect we are describing would also hold, in the sense that higher-level speakers (and not just the level-1 speaker) also prefer the message whose literal meaning is the most surprising. We choose to keep the model as simple as possible, to make the logic of our proposal maximally intelligible (cf. footnotes 16 and 4)

2.2 Choosing between I and O

Is the prediction in (11) correct?

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 (12a) below seems much more appropriate than using (12b).

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

Correspondingly, (13b) seems much more appropriate than (13a) in the same situation:

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

This is entirely in line with an explanation based on probabilistic expectations. In the specified context, the prior probability of the proposition that all talks will be given in English is higher than that of the proposition that no talk will be given in English. Correspondingly, the prior probability of (12a) (resp. (13b)) is smaller than that of (12b) (resp. (13a)), which predicts a preference for (12a) (resp. (13b)).

As observed by Roni Katzir (p.c.), one independent reason for the felicity contrasts in (13) and (12) might be due to the fact that *some* triggers a *not-many*-implicature, while *not every* might trigger a *many*-implicature. That is, e.g., (12b) might be interpreted as *some talks will be in English, but not many of them*, which in the situation we described might not be what the speaker intends to convey. And (13b) would be interpreted as conveying *some but not many of the talks will be in French*, which, on the contrary, might be exactly what the speaker wants to say. If the choice is between these two meanings, then only if the speaker happens to believe that many talks will be in French would she choose (13a), and this might play a role in the observed contrasts. However, we don't think the contrasts are markedly different if *some* is replaced with *many*. Suppose we are again in a context where it's highly expected (though not certain) that all talks will be in English, and we compare the following two sentences:

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

It seems to us that in such a context, one would be much more surprised to hear (14b) than (14a). Yet if both sentences in (14a) are strengthened into 'many but not all talks will be in English', we should not observe such an effect. The contrast is, however, fully expected in our account, simply because (14a) expresses a proposition whose prior probability, in the specified context, is smaller than that of (14b).

2.3 Is I most often more informative than O?

Is it in fact the case that, in most contexts, the condition stated in (10) ($P_0(w_\forall) < P_0(w_{\neg\exists})$) holds? Even though this is very hard to assess based on actual data, there are good reasons to think that it holds, as already discussed in Chater and Oaksford (1999). These authors observe, first, that the properties denoted by nouns, verbs and adjectives typically hold of a minority of objects (they call this observation the 'rarity assumption'): there are less cats than non-cats, and less red things than non-red things and presumably most often there

are less people who are singing than people who aren't. Note also that vague gradable adjectives like *tall* are typically interpreted in such a way that a minority of individuals within a comparison class count as tall (see Kennedy 2007 for discussion, among others). There are of course obvious counterexamples (*thing*, *exist*, ...), but overall, for most lexical predicates B, fewer things have the property B than the property non-B.[7] Second, Chater & Oaksford observed is that if predicates tend to be true of a small number of objects (as seems to be the case), then if we pick two predicates A and B randomly, we are much more likely to find that their intersection is empty than to find that A and B intersect (i.e. $P_0(s_\exists) < P_0(w_{\neg\exists})$), which entails the condition in (10).

In practice, however, the predicates A and B used in sentences of the form 'Q As are Bs' tend to be related in terms of their general subject matter. Typically, A denotes some small, cohesive region of the conceptual space (such as a species of animals, a nationality, a trade, the guests at a specific party, the talks at a specific conference, etc.), and B denotes a property that is well defined for A-objects (being of a certain color, doing a certain kind of activity, etc.). In these situations, the probability that No A is a B ($P_0(w_{\neg\exists})$) is not necessarily smaller than the probability that some As are B ($P_0(s_\exists)$). However, to the extent that, on most occasions, a randomly picked A is still more likely not to have property B than to have it (because predicates tend to denote minorities within a natural class), it will still be the case that it is less likely for all As to have property B than it is for all As not to have property B, which is exactly the condition stated in (10) ($P_0(w_{\neg\exists}) > P_0(w_\forall)$) Importantly, this condition is *weaker* than Chater and Oaksford's assumption that $P_0(s_\exists) < P_0(w_{\neg\exists})$, and is entailed by it. In Appendix B, we provide a model which explains why a lexicon where lexical predicates B typically apply to a minority of objects within a natural class A is optimal from an information-theoretic point of view. From this reasoning, we expect I statements to be more informative than O statements in a majority of situations where both kinds of statements are true.[8]

3 A model of the expected utility of a lexicon

The link between frequency of use and lexicalization can in principle be due to at least two types of pressures. One type of pressure is learnability: the more an expression is used, the easier it is to memorize it as a unit (see Hendrickson and Perfors 2019 for relevant discussion). The idea would be that, starting from a lexicon where the four corners are lexicalized, it will be easier for children to remember I than O, because they will hear I more often than O. As a result, such a language would be more likely to lose O than I when it is transmitted to the next generation. An explanation of this sort seems somewhat

[7] There are several reasons why this could be true. One is that 'natural' concepts typically cover a *connected* and relatively *homogeneous* region of the space of possible concepts (Gärdenfors 2004). To give an example, the *dog*-concept is arguably a more natural concept than the *non-dog* concept, because the concept of 'non-dog' includes many different types of objects which are intuitively extremely different from each other. This is the case even if we restrict our attention to a "natural" super-class of dogs, such as land animals or pets. In Appendix B, we discuss a further reason why the 'rarity' assumption may hold, based on information-theoretic considerations.

[8] The fact that a *some*-statement is on average more informative than a *not-all*-statement is not sufficient by itself to explain why the former is lexicalized while the latter is not. Very informative messages are by definition true in fewer situations than less informative ones, so even though they are particularly useful when they are used, there are also fewer situations where they can be used. We certainly do not want to predict that in general messages corresponding to unlikely events or situations are more likely to be lexicalized!

dubious in this specific case, since O-statements (expressed by 'not all' in English), though (as we argued) rarer than I-statements, are still not extremely rare. But there is in any case another well-known route to the same result (Zipf 1935; Piantadosi, Tily, and Gibson 2011, among many others): a language that lexicalizes frequent meanings as opposed to rare ones minimizes the average communicative effort of speakers (and parsing effort of listeners), compared to one where infrequent meanings, but not frequent meanings, would be lexicalized (in such a language, the meanings that you want to express most often require more words, and greater syntactic complexity). Conversely, speakers who seek to minimize their effort are likely to prefer inaccurate, simple expressions to accurate, complex ones; in a language where frequent meanings are lexicalized and therefore simple to express, this situation will be rarer and the average quality of information exchange will be higher.

In this section, we will offer a model of the *expected utility* of a lexicon in which a notion of cost is introduced. In any given situation, the model will assign a utility to each message, which will depend both on the cost of the message and its informativity (measured in relation with the prior probability distribution that characterizes the situation), and speakers are assumed to select the message that has the highest utility. We will be able to compute the *expected utility* of different lexica across occasions of uses. We will compare the expected utility of the attested lexicon $\{A, E, I\}$ with that of the unattested one, $\{A, E, O\}$. This approach formalizes the intuition that different lexicalization choices might lead to different outcomes in average communicative efficiency. In this section we offer a semi-formal account which contains the gist of our model (the fully explicit model is presented in Appendix A).

We aim to compare the expected utility of two languages. In the first language, the quantifiers A, E and I are lexicalized, while the quantifier O is expressed by a syntactically complex expression, while in the second one, A, E, and O are lexicalized and I is expressed by a complex expression. We summarize this in (15), where the superscript $^+$ signals that a quantifier is more complex than all others, which will translate into a specific *cost* :

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

We take the messages without the superscript to have a null cost, and the one with a superscript to have a positive cost c.

A situation of utterance consists of a pair $\langle w, P_0 \rangle$, where w is the actual world, by hypothesis known to the speaker, and P_0 is the probability distribution over worlds, corresponding to the beliefs of the listener in that situation (and that of the speaker before the speaker came to know w). As in RSA models that include message costs (Bergen, Levy, and Goodman 2016), the utility of a message m in a situation $\langle w, P_0 \rangle$ is given by:

(16) $U(m, w, P_0) = \log(P_0(w|[\![m]\!])) - \text{cost}(m)$.

Now, as discussed above, when w is either $\forall$ or $w_{\neg\exists}$, the message with the greatest utility is, in each language, A or E, since these messages have a null cost and are maximally informative. The comparison of the expected utility of each language thus hinges on what happens in the $\exists\neg\forall$-world. Let m be the 'cheap' message in the language under consideration (so $m = I$ in $\mathcal{M}_I$ and $m = O$ in $\mathcal{M}_O$) and m^+ be the expensive one (so $m^+ = O^+$ in $\mathcal{M}_I$ and $m^+ = I^+$ in $\mathcal{M}_O$).

We have:

(17) a. $U(m, w_{\exists\neg\forall}, P_0) = \log(P_0(w_{\exists\neg\forall}|[\![m]\!])$
b. $U(m^+, w_{\exists\neg\forall}, P_0) = \log(P_0(w_{\exists\neg\forall}|[\![m^+]\!]) - c$

Now:

(18) $P_0(w_{\exists\neg\forall}|[\![m]\!]) = \frac{P_0(w_{\exists\neg\forall}\cap[\![m]\!])}{P_0([\![m]\!])} = \frac{P_0(w_{\exists\neg\forall})}{P_0([\![m]\!])}$.[9]

Likewise,

(19) $P_0(w_{\exists\neg\forall}|[\![m^+]\!]) = \frac{P_0(w_{\exists\neg\forall})}{P_0([\![m^+]\!])}$

Therefore,

(20) a. $U(m, w_{\exists\neg\forall}, P_0) = \log(P_0(w_{\exists\neg\forall}|[\![m]\!]) = \log\frac{P_0(w_{\exists\neg\forall})}{P_0([\![m]\!])} = \log(P_0(w_{\exists\neg\forall}))-\log(P_0([\![m]\!]))$
b. $U(m^+, w_{\exists\neg\forall}, P_0) = \log(P_0(w_{\exists\neg\forall})) - \log(P_0([\![m^+]\!])) - c$

We assume that the speaker will pick the message with the greatest utility. That is, the speaker will use m^+ if and only if $U(m^+, w_{\exists\neg\forall}, P_0) > U(m, w_{\exists\neg\forall}, P_0)$, i.e., given the equalities above, whenever the following condition holds:

(21) $\log(P_0([\![m]\!])) - \log(P_0([\![m^+]\!])) > c$

The *informativity* of a proposition ϕ relative to P_0, noted $\text{Info}_{P_0}(\phi)$ is defined (in information theory) as $-\log(P_0(\phi))$. That is, the less probable the proposition expressed by a message is, the more *surprising* and informative it is. So we can rephrase the above condition as:

(22) $\text{Info}_{P_0}([\![m^+]\!]) - \text{Info}_{P_0}([\![m]\!]) > c$.

This makes sense: it says that the speaker will choose the more costly message just in case the gain in information relative to the cheaper message exceeds the extra cost of the more costly message (in case she believes both messages).

When applied to each language, this gives us:

(23) a. In $\mathcal{M}_I$, the costly message (O^+), which means *not all*, is used (in a situation where the speaker believes the world is $w_{\exists\neg\forall}$) if $\log(P_0(s_\exists))-\log(P_0(s_{\neg\forall})) > c$, i.e if $\text{Info}_{P_0}(s_{\neg\forall}) - \text{Info}_{P_0}(s_\exists) > c$; otherwise the cheap message I is used.
b. In $\mathcal{M}_O$, the costly message (I^+), which means *some*, is used if $\log(P_0(s_{\neg\forall})) - \log(P_0(s_\exists)) > c$, i.e. if $\text{Info}_{P_0}(s_\exists) - \text{Info}_{P_0}(s_{\neg\forall}) > c$; otherwise the cheap message O is used.

Now, in order to reason about expected utility, we need to take into account the fact that P_0 is not constant, i.e. varies across conditions of use and choices of predicates (technically, this means that P_0 is itself a random variable). We assume, following our discussion in section 2.3, that while P_0 varies, it is more often the case that $P_0(s_\exists) < P(s_{\neg\forall})$ than the reverse (see Appendix A for a precise way of expressing this assumption). Thus, the difference in informativity between I/I^+ and O/O^+, call it $Q = \text{Info}_{P_0}(s_\exists) - \text{Info}_{P_0}(s_{\neg\forall})$, will have a probability distribution across situations that is biased towards positive values. Figure 1 illustrates what such a distribution might look like. Focusing again on the case where the speaker believes $w_{\exists\neg\forall}$, we can essentially distinguish between two types of situations.

[9] Since m is either O or I, $w_{\exists\neg\forall}\cap[\![m]\!]$ reduces to $w_{\exists\neg\forall}$.

1. **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 $s_\exists$ is sufficiently low compared to that of $s_{\neg\forall}$ (so that the costly message, which means $s_\exists$, will be highly informative), namely when $\text{Info}_{P_0}(s_\exists) - \text{Info}_{P_0}(s_{\neg\forall}) > c$. This corresponds to the right-hand region on Figure 1. Meanwhile, in $\mathcal{M}_I$, this will happen when the prior probability of $s_{\neg\forall}$ is sufficiently low (so that the costly message, which means $s_{\neg\forall}$, is highly informative): this corresponds to the left-hand region on Figure 1, where $\text{Info}_{P_0}(w_\forall) - \text{Info}_{P_0}(s_\exists) > c$. Given our assumption that the first situation happens more often than the second one, we will use more often the costly message in $\mathcal{M}_O$ than in $\mathcal{M}_I$ — on Figure 1, this corresponds to the fact that the right-hand region has a larger area under the curve than the left-hand region. This creates a disadvantage for $\mathcal{M}_O$ on the cost side.

2. **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, where $|\text{Info}_{P_0}([\![m^+]\!]) - \text{Info}_{P_0}([\![m]\!])| < c$ (the absolute difference in informativity between the two messages is smaller than their difference in cost). In this case, the comparison of the utilities achieved by the two languages hinges on how informative the simpler message is relative to P_0: when P_0 is such that $P_0(s_\exists) < P_0(s_\neg\forall)$, 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 $P_0(s_\exists) > P_0(s_{\neg\forall})$, 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 the first situation ($P_0(s_\exists) < P_0(s_{\neg\forall})$) 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 corresponds to the fact that the right-hand half of the middle area is larger than the left-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.

This informal reasoning suggests that on average, the speaker of $\mathcal{M}_I$ will receive a higher utility than that of $\mathcal{M}_O$, both because she will use the costly message less often, and because when both types of speakers use the cheaper message, the speaker of $\mathcal{M}_I$ will be, most often, more informative than the speaker of $\mathcal{M}_O$.

In Appendix A, we provide a formally explicit model which captures this reasoning.

4 Discussion

4.1 Summary

Let us take stock.

In section 2 we suggested the following explanation for the non-lexicalization of O. English has a word for bakers but no word for people who don't sell bread, it has a word

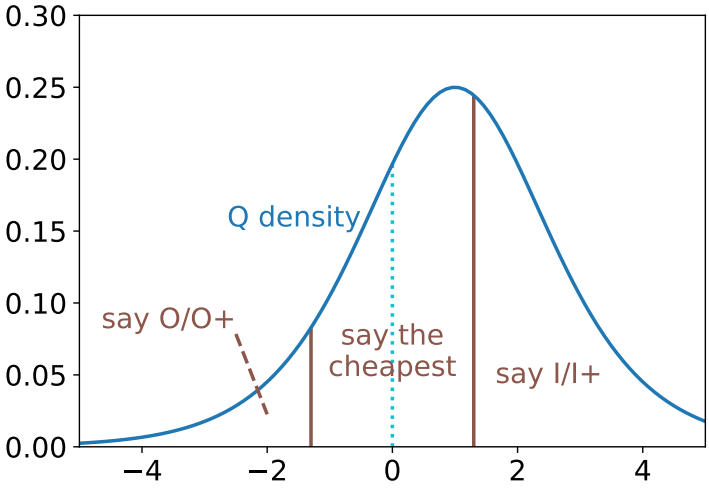

Figure 1: An example of a density for $Q = \log(P_0(s_{\neg\forall})) - \log(P_0(s_{\exists}))$ $(= \text{Info}_{P_0}(s_{\exists}) - \text{Info}_{P_0}(s_{\neg\forall}))$, biased towards larger values — which captures the fact that most often I-statements are more informative than O-statements. The x-dimension represents possible values of Q; it is divided into three intervals. In the first interval, speakers of both languages say O or O^+, because the meaning it expresses $(s_{\neg\forall})$ is much more informative than that expressed by I or I^+ $(s_{\exists})$; in the central interval, speakers of both languages say the cheaper message, because the difference in informativity (measured by Q) between the two messages is smaller than their difference in cost; in the third one, they say I or I^+, because its meaning $(s_{\exists})$ is much more informative than the meaning of O or O^+ $(s_{\neg\forall})$. The probability of each of these three cases is given by the area under the curve. The dotted blue line marks the limit between the domain where O, O^+ are more informative than I, I^+ and the one where the reverse is true.

for dogs but no word for non-dog animals, and so on. It follows that, on average, an O-statement is less informative than an I-statement, so that in situations where both types of statements are true, speakers will most often use I. Given that less frequent meanings tend to be lexicalized less than frequent meanings, it is to be expected that O will not be lexicalized, across languages, to the same extent as I is.

In Section 3, we have offered a more explicit approach where we compare the overall expected utility of different lexica, showing that a lexicon based on $\{A, E, I\}$ has a higher expected utility than one based on $\{A, E, O\}$.

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.[10]

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.[11]

[10] In the model discussed in section 3, they do so when O-statements are significantly more informative than I-statements, which will happen from time to time.

[11] 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 $w_{\exists\neg\forall}$ 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

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.

4.2 Comparison with other approaches

A strain of recent works have offered information-theoretic accounts for a number of properties of the lexicon of natural languages, based on the idea that languages maximize communicative efficiency through a trade-off between informativity and some notion of lexical complexity. This approach is most easily applied to content words, for which we can model the meaning space and the prior probabilities independently of linguistic facts. Examples include colour words (Zaslavsky, Kemp, et al. 2018), kinship terms (Kemp and Regier 2012), and animal names (Zaslavsky, Regier, et al. 2019). The main challenge that any attempt to extend the idea to logical vocabulary faces is that it needs to define notions of informativity and complexity over the abstract domains that it discusses, and how to do so is not entirely straightforward.

Steinert-Threlkeld (2020), Denić, Steinert-Threlkeld, and Szymanik (2021), and Uegaki (2020) are among the works that take up this challenge. All three of these works show that the attested lexica of quantifiers (Steinert-Threlkeld 2020; Denić, Steinert-Threlkeld, and Szymanik 2021) or connectives (Uegaki 2020) in Natural Language tend to perform better than unattested lexica on a certain metric of efficiency, and Uegaki (2020) specifically points out that this fact can be seen as a solution to Horn's puzzle. They adopt specific notions of informativity and complexity for messages.

On the informativity side, some variant of expected listener surprisal, similar to what we ourselves use, is universally adopted. This requires a notion of prior probability on possible situations. Steinert-Threlkeld (2020) and Uegaki (2020) choose to represent situations as (some description of) possible worlds and to assign the same probability to each world. Thus, it is impossible for them to demonstrate any effect of "real-world" distributions. The flat prior also makes it so that informativity alone cannot distinguish upwards and downwards monotone quantifiers, in the sense that I and O (and A and E as well) will necessarily be equally informative, because they are true in exactly the same number of worlds. The solution to Horn's puzzle will therefore have to come from the

essentially the line of argument of Horn (1973, pp. 251–260). However, as we pointed out, O statements are attested and are not used in the same contexts as I statements (cf. our discussion of examples (12) and (13) in section 2); for this reason, we are reluctant to just "shave off [the O category] with Occam's razor" (Horn 1973, p. 259).

complexity side. The way both Steinert-Threlkeld (2020) and Uegaki (2020) determine complexity is that they adopt a specific logical language based on common mathematical notation, including for instance the Boolean operators $\wedge$, $\vee$ and $\neg$, and they take a message's complexity to be the length of the shortest formula that represents it. As a consequence of this choice, downwards monotone operators are assumed from the get-go to be more complex than upwards monotone ones. We therefore expect lexica including O over I to be dispreferred, which is in fact what Uegaki (2020) finds. Thus, as a solution to Horn's puzzle, this approach is very similar to Horn's (1973) original proposal that there is an inherent semantic markedness to negation, and does not offer in turn an explanation for the latter fact (in this respect it is also similar to Katzir and Singh 2013). Where it improves on Horn's hypothesis is that it makes more specific predictions: Uegaki (2020) is able to explore the entire space of possible connective vocabulary.

An alternative to formulating arbitrary hypotheses about prior distributions and complexity is to derive them from linguistic data. This is what Denić, Steinert-Threlkeld, and Szymanik (2021) do. They adopt a classification of indefinites as well as a feature-based analysis due to Haspelmath (2001). They can then define the complexity of a message as the size of the smallest feature bundle that characterizes it. As far as the prior probability of each category is concerned, they estimate it from corpus data. Denić, Steinert-Threlkeld, and Szymanik (2021) do not discuss Horn's puzzle, and since they are exclusively concerned with indefinites, their model does not allow for a message meaning O.[12] We can still ask whether their method could offer a solution, if we were to adopt a similar classification of quantifiers (or connectives) and we found that inventories that satisfy Horn's generalization are more optimal. The main potential issue is one we already raised in Section 2. In principle, we want to derive informativity from the actual probability that a message is true. When we estimate the distribution of messages from corpus data, we are looking instead at the probability that a message is *produced*. If we consider that speakers take into account considerations of cost and complexity in their production, as in the model of Section 2.1, then what we are measuring already reflects the effects of cost and of the vocabulary of the language. Thus, if we find that positive existentials (I) are more common than negated universals (O), it might be that this is because they are easier to express, and not the other way around. There is a parallel concern on the complexity side: if we base our representation of messages on morphological patterns or cross-linguistic lexicalization patterns, then we are going to assign a more complex representation to O from the beginning. These two biases will make it so that our model already internalizes Horn's observation and cannot be used as an explanation for it.

The conclusion of this discussion is that existing information-theoretic approaches to the logical vocabulary of languages would not offer a complete explanation for Horn's puzzle, because the models they use already internalize in some form either Horn's observation that O is uncommon, or Horn's hypothesis that negation is marked.[13] In contrast,

[12] It should be noted that in saying that "languages lexicalize I," we have been abstracting away from the fact that most languages have a number of expressions with existential meaning, e.g. English *some*, *a*, *a certain*, *any*, *some ... or other*, *whichever* etc.

[13] In addition to the works we discussed, a somewhat different route is taken by Steinert-Threlkeld and Szymanik (2019), who show that certain semantic universals pertaining to quantifiers, such as permutation invariance, make quantifiers easier to learn by a Neural Network model. This could suggest that those universals emerge as an effect of learnability pressure. In this approach the learning properties of the network determine a notion of fitness that does not involve specific representational choices, beyond the architecture of the network itself. As in the other studies we discuss however, the observations that are presented to the model are drawn from an arbitrary distribution, so that no effect stemming from

we have derived the difference between the attested and the unattested lexicon entirely from the truth conditions of the messages, including a specific assumption about the prior probabilities of messages being true (as opposed to the probabilities of messages being used). We have also been able to derive the result analytically while leaving our assumptions somewhat abstract, e.g. without assuming particular probability distributions or particular costs. The price we pay is the extreme specifity of our result: as we have already noted, we are unable to extend the comparison beyond the two lexica that we discuss, and we do not account for any sort of pressure on lexicon size.

Our hope is that, despite the limitations we have just pointed out, and beyond the issue of Horn's puzzle, our work can serve as a further illustration of how explicit decision-theoretic models of pragmatics can in principle account, through the notion of expected utility, for certain universal tendencies in the logical lexicon of natural languages. Additionally, we hope to have shown that information-theoretic models can be used in linguistic research not just for broad, data-based analyses, but also to derive analytically specific qualitative points, in the same way as traditional formal analyses.

5 Acknowledgements

The research leading to these results has received funding from the Agence Nationale de la Recherche (FrontCog Project ANR-17-EURE-0017 and ProbaSem project ANR-19-CE28-0004-01). The authors thank Emmanuel Chemla and Isabelle Dautriche for useful discussions.

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.

Buccola, Brian and Benjamin Spector (2016). "Modified numerals and maximality". In: *Linguistics and philosophy* 39.3, pp. 151–199.

Chater, Nick and Mike Oaksford (1999). "The probability heuristics model of syllogistic reasoning". In: *Cognitive psychology* 38.2, pp. 191–258.

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

Denić, Milica, Shane Steinert-Threlkeld, and Jakub Szymanik (2021). "Complexity/informativeness trade-off in the domain of indefinite pronouns". In: *Proceedings of SALT 30*. To appear.

Franke, Michael (2011). "Quantity implicatures, exhaustive interpretation, and rational conversation". In: *Semantics and Pragmatics* 4.1, pp. 1–82.

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

real-world prior distributions can be demonstrated; furthermore, formal properties of the model and the mode of presentation of the data (consisting in binary encodings of Kripke-style models) make it so that flipping downwards and upwards monotone quantifiers (such as flipping I and O, or A and E) cannot affect the results. Thus the model offers no insight towards Horn's puzzle.

Geurts, Bart and Frans van Der Slik (2005). "Monotonicity and processing load". In: *Journal of semantics* 22.1, pp. 97–117.
Gibson, Edward et al. (2019). "How efficiency shapes human language". In: *Trends in cognitive sciences* 23.5, pp. 389–407.
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.
Haspelmath, Martin (2001). *Indefinite pronouns*. Oxford University Press.
Hendrickson, Andrew T and Andrew Perfors (2019). "Cross-situational learning in a Zipfian environment". In: *Cognition* 189, pp. 11–22.
Horn, Laurence Robert (1973). "On the semantic properties of logical operators in English". Ph.D. dissertation, UCLA.
Katzir, Roni, Nur Lan, and Noa Peled (2020). "A note on the representation and learning of quantificational determiners". In: *Proceedings of Sinn und Bedeutung*. Vol. 24. 1, pp. 392–410.
Katzir, Roni and Raj Singh (2013). "Constraints on the lexicalization of logical operators". In: *Linguistics and Philosophy* 36.1, pp. 1–29.
Kemp, Charles and Terry Regier (2012). "Kinship categories across languages reflect general communicative principles". In: *Science* 336.6084, pp. 1049–1054.
Kennedy, Christopher (2007). "Vagueness and grammar: The semantics of relative and absolute gradable adjectives". In: *Linguistics and philosophy* 30.1, pp. 1–45.
Penka, Doris (2011). *Negative indefinites*. 32. Oxford University Press Mexico SA De CV.
Piantadosi, Steven T., Harry Tily, and Edward Gibson (2011). "Word lengths are optimized for efficient communication". In: *Proceedings of the National Academy of Sciences* 108.9, pp. 3526–3529. eprint: https://www.pnas.org/content/108/9/3526.full.pdf. URL: https://www.pnas.org/content/108/9/3526.
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.
Steinert-Threlkeld, Shane (2020). "Quantifiers in natural language optimize the simplicity/informativeness trade-off". In: *Proceedings of the 22nd Amsterdam colloquium*, pp. 513–522.
Steinert-Threlkeld, Shane and Jakub Szymanik (2019). "Learnability and semantic universals". In: *Semantics and Pragmatics* 12, p. 4.
Uegaki, Wataru (2020). "*NAND and the communicative efficiency model". Ms. University of Edinburgh.
Zaslavsky, Noga, Charles Kemp, et al. (2018). "Efficient compression in color naming and its evolution". In: *Proceedings of the National Academy of Sciences* 115.31, pp. 7937–7942. eprint: https://www.pnas.org/content/115/31/7937.full.pdf. URL: https://www.pnas.org/content/115/31/7937.
Zaslavsky, Noga, Terry Regier, et al. (2019). "Semantic categories of artifacts and animals reflect efficient coding". In: *41st Annual Conference of the Cognitive Science Society*.
Zeijlstra, Hedde (2011). "On the syntactically complex status of negative indefinites". In: *The Journal of Comparative Germanic Linguistics* 14.2, pp. 111–138.
Zipf, GK (1935). *The psychobiology of language*. Houghton, Mifflin.

Appendix A Comparing the utilities of two lexica

Preliminaries In order to make the reasoning presented in section 3 fully explicit, we need to make part of our model the fact that P_0, the prior distribution, is not given, and instead varies across situations. We can do this by making P_0 a random variable of its own, which ranges over possible probability distributions over the universe Ω (recall that Ω is the 3-element set $\{w_{\neg\exists}, w_{\exists\neg\forall}, w_{\forall}\}$).

This new assumption allows us to define the *expected utility* of a lexicon. For a lexicon $\mathcal{M}$ and a prior P_0, let us denote as $m(P_0, \mathcal{M})$ the message that a speaker of $\mathcal{M}$ will use in the situation $\langle w_{\exists\neg\forall}; P_0\rangle$ (when the world is $w_{\exists\neg\forall}$ and the prior is P_0). The utility achieved by the speaker in this situation can be written as $U(m(P_0, \mathcal{M}), w_{\exists\neg\forall}, P_0)$. Taking the expected value of this quantity over all potential prior distributions P_0, we obtain what we can call the conditional expected value of the lexicon,[14] which we write as $\bar{U}(\mathcal{M})$ and which is given in (24). This quantity represents the average utility achieved by the speaker of $\mathcal{M}$ in $w_{\exists\neg\forall}$-situations.

(24) $\bar{U}(\mathcal{M}) = \mathbb{E}[U(m(P_0, \mathcal{M}), w_{\exists\neg\forall}, P_0)]$

$\bar{U}(\mathcal{M})$ is what we will use to compare $\mathcal{M}_I$ and $\mathcal{M}_O$ in a formal way: what we want to derive is that $\bar{U}(\mathcal{M}_I) > \bar{U}(\mathcal{M}_O)$, corresponding to the idea that speakers of $\mathcal{M}_I$ achieve greater utility on average.

We also need to formalize the idea that P_0 is most often such that $P_0(w_{\neg\exists}) > P_0(w_{\forall})$, that is, that in most situations speakers consider that no As being Bs is more likely than all As being Bs. There are probably various ways this could be done. Here is what we are going to assume: for any particular distribution P_0' which is biased in favor of $w_{\forall}$ relative to $w_{\neg\exists}$, we assume that the distribution P_0'' which encodes a bias of the same magnitude in the opposite direction is more likely to be the one that characterizes the listener's epistemic state. That is, for a particular distribution P_0', consider P_0'' which is like P_0', except that the probabilities of $w_{\neg\exists}$ and $w_{\forall}$ are flipped:

(25)
$$\begin{aligned} P_0''(w_{\neg\exists}) &= P_0'(w_{\forall}) \\ P_0''(w_{\exists\neg\forall}) &= P_0'(w_{\exists\neg\forall}) \\ P_0''(w_{\forall}) &= P_0'(w_{\neg\exists}) \end{aligned}$$

If $P_0'(w_{\neg\exists}) > P_0'(w_{\forall})$, then $P_0''(w_{\neg\exists}) < P_0''(w_{\forall})$, and *vice-versa*. In other words, at most one of P_0' and P_0'' is such that the condition we expect to be the most common case is true. We are going to assume that this one is more likely to be the actual P_0 than the other:

(26) **Bias assumption (BA):** if Φ is the density of the variable P_0, and if P_0' and P_0'' are related in the way described above, then:
   a. If $P_0'(w_{\neg\exists}) > P_0'(w_{\forall})$, then $\Phi(P_0') > \Phi(P_0'')$.
   b. If $P_0'(w_{\neg\exists}) < P_0'(w_{\forall})$, then $\Phi(P_0') < \Phi(P_0'')$.

[14] What makes $\bar{U}(\mathcal{M})$ "conditional" is that we only consider what happens in $w_{\exists\neg\forall}$-situations. It would be natural to define "expected utility" as the expectation of the utility achieved taken across all possible situations. $\bar{U}(\mathcal{M})$ is what we get if we condition on the fact that the world is $w_{\exists\neg\forall}$. The reason that we consider $\bar{U}(\mathcal{M})$ and not "proper" expected utility here is that the two languages we want to compare achieve exactly the same utility in $\neg\exists$-situations as well as $\forall$-situations, and therefore these situations will not matter to the comparison. In other words, whichever language yields a greater conditional expected utility also yields a greater expected utility.

The assumption in (26) is how we will capture the fact that most of the time $P_0(w_{\neg\exists}) > P_0(w_\forall)$. If a particular choice of P_0 does not respect the condition, we assume that it is less likely than its mirror image that does respect it.

The BA is strictly stronger than our initial statement that "$P_0(w_{\neg\exists}) > P_0(w_\forall)$ is usually true", which one would most naturally implement as (27). (27) is in fact insufficient to derive the desired result: it might be that (27) is true, and that $\bar{U}(\mathcal{M}_I) < \bar{U}(\mathcal{M}_O)$. This will be the case for instance if the most probable values of P_0 are either such that $P_0(w_{\neg\exists})$ is much smaller than $P_0(w_\forall)$, or such that $P_0(w_{\neg\exists})$ is slightly greater than $P_0(w_\forall)$. However, we think that there is no reason why the distribution of P_0 should exhibit such an asymmetrical shape. The distributions commonly used in mathematical modeling usually have simple shapes that are obtained from smooth deformations of perfectly symmetrical ones, with a single maximum or minimum. While real-world data can of course depart from this pattern, it is usually due to a well-identified categorical effect and we see no reason to think that it should happen in this instance. If we restrict ourselves to common parametric families, it is in fact the case that (27) and the BA are equivalent; in other words, any natural choice of parametrization for P_0 is such that the assumption in (27) would entail the BA.[15] For this reason, we do not think that our implementation of the bias assumption affects the generality of our result.

(27) $\quad \mathbb{P}(P_0(w_{\neg\exists}) > P_0(w_\forall)) > \frac{1}{2}$

The proof We can now prove the desired result: if the bias assumption holds, then $\bar{U}(\mathcal{M}_I) > \bar{U}(\mathcal{M}_O)$.

In section 3, we derived the following results:

(28)
a. If $\log(P_0(s_\exists)) - \log(P_0(s_{\neg\forall})) > c$, then $m(P_0, \mathcal{M}_I) = O^+$ and $m(P_0, \mathcal{M}_O) = O$.
b. If $\log(P_0(s_{\neg\forall})) - \log(P_0(s_\exists)) > c$, then $m(P_0, \mathcal{M}_I) = I$ and $m(P_0, \mathcal{M}_O) = I^+$.
c. If neither condition holds, then $m(P_0, \mathcal{M}_I) = I$ and $m(P_0, \mathcal{M}_O) = O$.

This can be rewritten in terms of the following quantity Q (Q being a function of the random variable P_0, it is also a random variable):

(29) $\quad Q := \log(P_0(s_{\neg\forall})) - \log(P_0(s_\exists))$

[15] We think that the most natural choice of a parametrization for a 3-way probability distribution like P_0 would be a Dirichlet distribution $\mathrm{Dir}(\alpha, \beta, \gamma)$. Once we adopt this parameterization, (27) and the BA are both equivalent to $\alpha > \gamma$, so that the BA is innocuous.

Another, even simpler way to parameterize P_0 is to make the simplifying assumption that, when considering a sentence of the form QAB, the prior probability that a given A-individual x has property B is independent of the probability that some other individual A-individual y has property B, and that this probability is uniform across As. Then, P_0 depends entirely on the parameter p_0, the probability that a given A is a B, and the number of A-individuals. We have: $P_0(w_\forall) = p_0^n$ and $P_0(w_{\neg\exists}) = (1-p_0)^n$, where n is the number of As. Thus the condition $P_0(w_{\neg\exists}) > P_0(w_\forall)$ is equivalent to $p_0 < 0.5$, and (27) is equivalent to the density of p_0 having more mass on the left-hand side of the graph. If p_0 follows a Beta law (as would be most natural for a Bernoulli parameter), the distribution has a simple shape (e.g. a bell shape in the case where the parameters are greater than 1) and (27) will be true if and only if the density function tilts to the left, which would make the BA true as well.

A consequence of these facts is that if we had demonstrated our point through numerical simulations, as is common in the literature applying information-theoretic models to linguistics, the distributions we would have looked at would have been such that our result would have followed from (27). Because we want to derive an analytical result instead, we are forced to make our assumptions explicit, but all in all this makes our result more general, not less.

(30) a. If $Q < -c$, then $m(P_0, \mathcal{M}_I) = O^+$ and $m(P_0, \mathcal{M}_O) = O$.
b. If $Q > c$, then $m(P_0, \mathcal{M}_I) = I$ and $m(P_0, \mathcal{M}_O) = I^+$.
c. If $|Q| < c$, then $m(P_0, \mathcal{M}_I) = I$ and $m(P_0, \mathcal{M}_O) = O$.

Note that Q is exactly the difference in informativity between I and O: $Q = \text{Info}([\![I]\!]) - \text{Info}([\![O]\!])$. This makes the above pattern intuitive: when I is much more informative, say I or I^+; when O is much more informative, say O or O^+; when the difference is small, say the cheapest.

What follows is a formal calculation that does not invoke any specific insight; the reader who is not interested in checking the correctness of it can skip straight to (36). We begin by decomposing $\bar{U}$ based on the value of Q:

$$\begin{aligned}(31)\qquad \bar{U}(\mathcal{M}_I) = {} & \mathbb{P}(Q < -c)\mathbb{E}[U(O^+, w_{\exists\neg\forall}, P_0) \mid Q < -c] \\ & + \mathbb{P}(Q > c)\mathbb{E}[U(I, w_{\exists\neg\forall}, P_0) \mid Q > c] \\ & + \mathbb{P}(|Q| < c)\mathbb{E}[U(I, w_{\exists\neg\forall}, P_0) \mid |Q| < c]\end{aligned}$$

Furthermore, for any condition C:

$$\begin{aligned}(32)\qquad \mathbb{E}[U(O^+, w_{\exists\neg\forall}, P_0) \mid C] &= \mathbb{E}[\log P_0(w_{\exists\neg\forall}) - \log P_0(s_{\neg\forall}) - c \mid C] \\ &= \mathbb{E}[\log P_0(w_{\exists\neg\forall}) \mid C] - \mathbb{E}[\log P_0(s_{\neg\forall}) \mid C] - c\end{aligned}$$

And similarly:

$$\begin{aligned}(33)\qquad \mathbb{E}[U(I, w_{\exists\neg\forall}, P_0) \mid C] &= \mathbb{E}[\log P_0(w_{\exists\neg\forall}) - \log P_0(s_{\exists}) \mid C] \\ &= \mathbb{E}[\log P_0(w_{\exists\neg\forall}) \mid C] - \mathbb{E}[\log P_0(s_{\exists}) \mid C]\end{aligned}$$

Putting these two together:

$$\begin{aligned}(34)\qquad \bar{U}(\mathcal{M}_I) = {} & \mathbb{P}(Q < -c)\mathbb{E}[\log P_0(w_{\exists\neg\forall}) \mid Q < -c] + \mathbb{P}(Q > c)\mathbb{E}[\log P_0(w_{\exists\neg\forall}) \mid Q > c] \\ & + \mathbb{P}(|Q| < c)\mathbb{E}[\log P_0(w_{\exists\neg\forall}) \mid |Q| < c] - c\mathbb{P}(Q < -c) \\ & - \mathbb{P}(Q < -c)\mathbb{E}[\log P_0(s_{\neg\forall}) \mid Q > -c] - \mathbb{P}(Q > c)\mathbb{E}[\log P_0(s_{\exists}) \mid Q > c] \\ & - \mathbb{P}(|Q| < c)\mathbb{E}[\log P_0(s_{\exists}) \mid |Q| < c]\end{aligned}$$

We can do the same thing with $\bar{U}(\mathcal{M}_O)$, and we derive:

$$\begin{aligned}(35)\qquad \bar{U}(\mathcal{M}_O) = {} & \mathbb{P}(Q < -c)\mathbb{E}[\log P_0(w_{\exists\neg\forall}) \mid Q < -c] + \mathbb{P}(Q > c)\mathbb{E}[\log P_0(w_{\exists\neg\forall}) \mid Q > c] \\ & + \mathbb{P}(|Q| < c)\mathbb{E}[\log P_0(w_{\exists\neg\forall}) \mid |Q| < c] - c\mathbb{P}(Q > c) \\ & - \mathbb{P}(Q < -c)\mathbb{E}[\log P_0(s_{\neg\forall}) \mid Q > -c] - \mathbb{P}(Q > c)\mathbb{E}[\log P_0(s_{\exists}) \mid Q > c] \\ & - \mathbb{P}(|Q| < c)\mathbb{E}[\log P_0(s_{\neg\forall}) \mid |Q| < c]\end{aligned}$$

When we take the difference, most terms cancel out:

$$\begin{aligned}(36)\qquad \bar{U}(\mathcal{M}_I) - \bar{U}(\mathcal{M}_O) &= c(\mathbb{P}(Q > c) - \mathbb{P}(Q < -c)) \\ &\quad + \mathbb{P}(|Q| < c)\mathbb{E}[\log P_0(s_{\neg\forall}) - \log P_0(s_{\exists}) \mid |Q| < c] \\ &= c(\mathbb{P}(Q > c) - \mathbb{P}(Q < -c)) + \mathbb{P}(|Q| < c)\mathbb{E}[Q \mid |Q| < c]\end{aligned}$$

The remaining terms can be given intuitive interpretations. $c(\mathbb{P}(Q > c) - \mathbb{P}(Q < -c))$ is the difference in expected cost between the two languages. Speakers of $\mathcal{M}_I$ use costly messages when $Q < -c$, while speakers of $\mathcal{M}_O$ use costly messages when $Q > c$. The other term is the difference in expected informativity. The two languages result in different

informativity only in situations where their speakers use the cheapest message, i.e. when $|Q| < c$; in these situations, as we have seen, Q quantifies the difference in informativity.

In section 3, we argued that both terms should be positive. This fact in fact follows from the bias assumption. To begin with, let us call ϕ the density of Q. It follows from the bias assumption that the following holds:

(37) For any $q > 0$, we have $\phi(q) > \phi(-q)$.

This is because mirroring P_0 as done per the bias assumption turns Q into $-Q$, and the variant that we assume to be more likely is also the one that yields a positive value for Q.

Then, we have:

$$\begin{aligned}(38)\qquad \mathbb{P}(Q > c) - \mathbb{P}(Q < -c) &= \int_c^{+\infty} \phi(q)\mathrm{d}q - \int_{-\infty}^{-c} \phi(q)\mathrm{d}q \\ &> \int_c^{+\infty} \phi(-q)\mathrm{d}q - \int_{-\infty}^{-c} \phi(q)\mathrm{d}q \quad \text{(per the BA)} \\ &> \int_{-\infty}^{-c} \phi(q)\mathrm{d}q - \int_{-\infty}^{-c} \phi(q)\mathrm{d}q \\ &> 0.\end{aligned}$$

And:

$$\begin{aligned}(39)\qquad \mathbb{E}[Q \,|\, |Q| < c] &= \frac{1}{\mathbb{P}(|Q| < c)} \int_{-c}^{c} \phi(q) q \mathrm{d}q \\ &= \frac{1}{\mathbb{P}(|Q| < c)} \left(\int_{-c}^{0} \phi(q) q \mathrm{d}q + \int_{0}^{c} \phi(q) q \mathrm{d}q \right) \\ &> \frac{1}{\mathbb{P}(|Q| < c)} \left(\int_{-c}^{0} \phi(q) q \mathrm{d}q + \int_{0}^{c} \phi(-q) q \mathrm{d}q \right) \quad \text{(per the BA)} \\ &> \frac{1}{\mathbb{P}(|Q| < c)} \left(\int_{-c}^{0} \phi(q) q \mathrm{d}q - \int_{-c}^{0} \phi(q) q \mathrm{d}q \right) \\ &> 0.\end{aligned}$$

It follows that $\bar{U}(\mathcal{M}_I) > \bar{U}(\mathcal{M}_O)$, as desired.[16]

Appendix B Expected utility and the optimality of predicates

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.[17] Having

[16] While we do not provide a full proof, this result still holds if we relax the assumption that speakers are perfectly rational in selecting their message, and we assume instead that they emit messages stochastically following a soft-max rule, as in the standard RSA model. This follows directly from the following fact, which holds both with the soft-max rule and in the model with full rationality: if P_0' and P_0'' are mirror images as in (25), then the expected utility achieved by speakers of $\mathcal{M}_I$ when $P_0 = P_0'$ is the same as the expected utility achieved by speakers of $\mathcal{M}_O$ when $P_0 = P_0''$, and vice-versa. The calculation would in fact be more direct in the general case, but we think that the version with full rationality, where one can clearly identify a cost term and an informativity term, is more enlightening.

[17] 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

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:

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

We ignore here the possibility of compositionally complex messages such as "A but not B".[18] Thus we assume there are two possible messages:

(41) A: "A!"
B: "B!"

Their semantics are the obvious ones:

(42) $[\![A]\!] = \Omega$
$[\![B]\!] = \text{B}$

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 that the prior distribution over world-states 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.

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

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 (43), 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

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.

[18] 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.

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. We do not predict through this reasoning that the prior of B should be particularly low, only that it should be below 0.5. Such a weak prediction is desirable given our discussion in section 2.3.[19]

[19]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. 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 (large) minority of a given comparison class.