Plausibility guides interpretation: Evidence from scope ambiguity resolution

Abstract

How do we navigate ambiguity in natural language? The current paper combines evidence from corpus analysis, behavioral experiments, and computational modeling to develop a formal model of the utterance-understanding process that yields disambiguation in context. Focusing on quantifier-negation sentences (e.g., *Every vote doesn't count*), we find evidence that listeners (i) evaluate the plausibility of competing interpretations, and (ii) are more likely to attribute to speakers interpretations that are more likely to be true, thereby aligning their interpretations with prior beliefs about the world. In this case study, plausibility is grounded in shared expectations about likely states of the world, allowing us to explain observed interpretation variation in naturalistic data.

Keywords: scope ambiguity, pragmatic reasoning, plausibility, corpus analysis, Rational Speech Act model

1 Introduction

Resolving ambiguity is a core challenge for human cognition. In natural language, ambiguity often arises through competing interpretations of an utterance. In fact, natural languages are "massively ambiguous" (Wasow et al., 2005, pp. 1) in that they are full of expressions with a form that maps to multiple potential meanings, especially when the expression is considered without context. Given all of this ambiguity, how do speakers manage to communicate successfully and what does context do to help listeners disambiguate? The current paper investigates one factor that can help listeners disambiguate utterances in context: the relative plausibility of competing interpretations, or how well an interpretation fits with a listener's prior beliefs about the world. Interpretations that are more likely to be true—that is, more plausible—are more likely to be selected.

We can see the practical problem ambiguity presents for human cognition when we consider the long-standing problem it presents for computer systems, which sometimes generate hundreds of not-obviously-invalid parses for a single natural language sentence (e.g., Manning and Schütze, 1999; Jusoh, 2018). Additionally, for humans, structural ambiguity is associated with increased online processing difficulty if a listener has to reanalyze a sentence because their initial parse seems incorrect (e.g., Frazier and Fodor, 1978; MacDonald et al., 1994). On the other hand, structural ambiguity is sometimes not at all associated with processing difficulty, with many studies showing no reading slowdown during ambiguity, and the evidence is generally mixed for whether difficulty arises within the syntactically ambiguous region itself (e.g, Clifton Jr and Staub, 2008; Grant et al., 2020). On a broader view, the intended meanings of most language expressions on their own are fundamentally under-determined, and listeners perform inferences in context in order to understand what the speaker meant to communicate (Grice, 1975; Sperber and Wilson, 1986).

In this paper, we investigate how ambiguity resolution could proceed when the ambiguous sentence is encountered as communication in context. We focus on sentence ambiguity, asking what interpretations are preferred for scopally-ambiguous utterances (e.g., *Every vote doesn't count*) and why certain interpretations may be preferred in certain contexts—where *context* is characterized by

prior expectations about the world. We show that prior expectations are indeed useful: listeners can rely on them to help disambiguate a speaker's meaning in favor of the more plausible interpretation.

To better understand scope ambiguity, suppose that the meaning of a phrase or sentence is like a formula successively built up of smaller formulae that combine in a particular order. The problem for natural language interpretation is that there are no parentheses in the overt form of the expression to signal the relative scope of its parts. A further problem is shown in (1): does *all* apply to *the glasses* and then to *are not clean* or does *not* apply to *all the glasses are clean*? In other words, when there are multiple logical operators (e.g., quantifying expressions like *all* and negating expressions like *not*), their order of operation might not reflect their surface order in the expression. For instance, in (1b), *not* applies first, although it is not said first. So, deciding the order of operations is an exercise in ambiguity resolution.

(1) All the glasses are not clean.
   a. ((all the glasses) are not clean) *paraphrased as* none of the glasses are clean
   b. (not (all the glasses are clean)) *paraphrased as* not all of the glasses are clean

Scope ambiguity has received much attention in previous theoretical and empirical work and has been studied from primarily semantic and syntactic perspectives, investigating whether and how ambiguity arises when there are multiple scope-taking operators in the same clause (e.g., Reinhart, 1983; May, 1985; Kurtzman and MacDonald, 1993; Musolino, 1999; Szabolcsi, 2011; Kiss and Pafel, 2017). Because the phenomenon is broad, including different constructions and scope-taking operators, we focus on the kind of scope ambiguity in (1), with a quantified subject (*all the glasses*) preceding sentential negation (*are not clean*)—hereafter, *quantifier-negation* utterances.

To illustrate our approach and preview some of our findings, consider the quantifier-negation utterance in (2), uttered in 2004 on a CNN segment (Davies, 2015). Like (1), (2) is potentially ambiguous between the two interpretations in (2a) and (2b), depending on the logical scope of the quantifier *every* relative to negation. Achieving the surface scope interpretation in (2a) involves the quantifier *every* taking scope over negation *not*, in line with their surface order in the utterance. This order leads to the interpretation that nothing under discussion has moved to California. In contrast, for the inverse scope interpretation in (2b), negation *not* takes scope over the quantifier *every* in inverse order to their use in the utterance. This configuration leads to the interpretation that it is not the case that everything under discussion has moved to California.

(2) Everything has not moved to California.
   a. Nothing has moved to California. — *surface scope* (every > not)
   b. Not all things have moved to California. — *inverse scope* (not > every)

Plausibility as determined by the context can strongly influence the preferred interpretation. Out of context, (2) may appear ambiguous and somewhat unusual. For instance, why would the speaker not use salient, unambiguous alternative phrases, such as *nothing* (leading to the surface scope interpretation) or *not everything* (leading to the inverse scope interpretation) has moved to California? Yet the original conversational context, shown in (3), is rich in information that motivates and disambiguates the utterance.

(3) CALLER: Hi. My question for Mr. Eisner was, MGM is one of my favorite places in Disneyworld and one of my favorite attractions there is the animation studios, and now the studio, the animation studio there is closed, and everything has moved to California, and I wanted to know how you justified doing that.
EISNER: Well, **everything has not moved to California**. We will still be demonstrating

animation in Florida.

The overall intuition we seek to quantitatively investigate, concerning the role of plausibility in disambiguation, is this: in (3), the fact that the first speaker believes that *all* animation studios have moved to California provides a cue that the subsequent use of *every*-negation was intended with its *not all* (inverse scope) reading rather than with its *none* (surface scope) reading. Specifically, the first speaker expresses the non-negated version of the subsequent *every*-negation use: *everything has moved to California*. This at least expresses the belief that the animation studios under discussion tend to move to California. So it is believed far more likely that *some or all* animation studios have moved than that *none* of the animation studios have moved (i.e., where $p(world)$ is the believed probability that a *world* state is true: $p(some, all) > p(none)$, and so $p(some, all) - p(none) > 0$). The greater this difference believed to hold between the *some or all* world states relative to the *none* world state (i.e., the greater $p(some, all) - p(none)$), the more the listener will reason that the speaker of (2) cannot have meant the *none* (surface) interpretation and, therefore, meant the *not all* (inverse) interpretation. In other words, because $p(none)$ is so low *a priori*, the *none* interpretation has a low probability of being true (i.e., low plausibility). One pressure (among many) on listeners' disambiguation is to align their interpretations with their *a priori* understanding of the world.

To formalize this intuition, we implement the hypothesized reasoning in context using a computational cognitive model formulated within the Bayesian Rational Speech Act (*RSA*) modeling framework (Frank and Goodman, 2012; Goodman and Frank, 2016). According to the hypothesis implemented via RSA, expectations about the true state of the world help to predict scope preference, because they make one scope interpretation relatively more plausible and therefore more probable. These expectations are salient to interlocutors and can even be expressed in the preceding linguistic context of the potentially-ambiguous utterance, as in (3).

Specifically, we assume boundedly rational speakers who try to minimize the cost of speaking while maximizing the probability that listeners arrive at the intended interpretation, given limits on the linguistic knowledge and information available to listeners (e.g., limits on experience with certain syntactic forms). In formalizing our hypothesis as a computational cognitive model, we specify prior expectations about the world that are skewed a particular way, affecting how plausible the different interpretations seem to the listener. We then analytically explore how this kind of context is useful for interpretation success and for capturing variation in interpretation preferences as observed in both naturalistic speech and controlled behavioral experiments. Upon hearing a quantifier-negation utterance, listeners assume that the speaker said something that is true and they reason that the speaker's intended interpretation is the one that is more likely to be true (i.e., the interpretation that is more plausible). We therefore predict that, given skewed priors, certain scope interpretations are more probable because they are more likely to be true.

We note that our goal is not to model the real-time processing of language, but rather the interpretation preferences that arise from it. At the same time, the RSA framework integrates multiple sources of information in a single probabilistic inference, and is therefore perhaps more naturally aligned with interactive approaches to comprehension than with strictly modular ones. No encapsulation is assumed in the RSA framework. We should also state here at the outset that our aim is to examine plausibility as one of the factors influencing interpretation preferences, not as the only factor. Multiple competing forces likely shape how listeners interpret quantifier-negation utterances (for discussion, see Scontras and Pearl, 2021). Our goal is to formalize and empirically evaluate our hypothesis about the role of interpretation plausibility in the pragmatics of scope disambiguation. By doing so, we hope the current investigation will set the stage for future work on additional factors and thereby facilitate a fuller understanding of utterance disambiguation. We return to this point at the paper's conclusion.

The remainder of this paper is organized as follows. Section 2 presents an overview of the literature and our theoretical foundations, highlighting how our reading of the literature suggests that the relative plausibility of competing interpretations is one reason for interpretation preferences. Section 3 formally specifies this hypothesis about the role of plausibility with a computational model, which describes the utility of world-expectations-as-skewed-priors for a listener who interprets *every*-negation utterances. We test analytically test how the model predictions about preferred interpretations depend on such expectations about the world. In Section 4, we investigate how *every*-negation utterances are used in everyday speech by collecting naturalistic examples from English corpora and crowd-sourced annotations of their preferred interpretations in context. To investigate the role of plausibility, we measure world expectations in the linguistic contexts of the corpus of *every*-negation utterances. We then test how well plausibility, as operationalized by the world expectations we measure, predicts average inverse scope preference per item and even per individual judgment. Finally, in Section 5, we turn to investigating the generalizability of the model. We use a controlled behavioral experiment to evaluate whether our utterance disambiguation model, originally formulated for *every*-negation utterances, can generalize to different cases of quantifier-negation scope ambiguity. Section 6 concludes by discussing how our findings relate to everyday ambiguity resolution, the value of studying naturalistic speech, implications for our understanding of scope ambiguity from a listener's perspective, and future directions for research.

2 The role of context and plausibility in interpretation preferences

The literature on quantifier-negation scope ambiguity in English tends to focus on *every*-negation and *all*-negation and, as a whole, finds variation in preferences for surface vs. inverse scope.[1] On the one hand, converging evidence from adults, children, and non-native English speakers suggests that surface scope is easier to access for any scopally-ambiguous utterance (Musolino, 1999; Musolino and Lidz, 2003; Viau et al., 2010; Lidz, 2018; Chung and Shin, 2022). This finding is mainly based on truth-value judgments, either spoken (e.g., Musolino, 1999) or written and in combination with self-paced reading (Chung and Shin, 2022). The finding is in line with surface scope being a general default in grammatical representations or processing (Lakoff, 1971; May, 1985; Pritchett and Whitman, 1995; Tunstall, 1998; Anderson, 2004; Scontras et al., 2017).

On the other hand, experimental studies show that adults prefer inverse scope interpretations of *every*- and *all*-negation utterances (Carden, 1970; Heringer, 1970; Carden, 1973; Musolino et al., 2000). These studies use a range of methodologies, including linguistic interviews (Carden, 1972; Musolino et al., 2000) and graded acceptability judgments of the written use of an utterance in context (Heringer, 1970). This preference for inverse scope also aligns with findings from corpus studies of English: Musolino et al. (2000) cite in a footnote that 28 out of 30 *every*-negation uses collected from English spontaneous speech were intended with inverse scope (their method of collecting scope judgments is unclear), and Neukom-Hermann (2016) finds that 54% of 469 *all*-negation uses from the British National Corpus (which is primarily written) were intended with inverse scope and only 17% with surface scope, as judged by Neukom-Hermann (the remainder were judged to have a third type of interpretation called collective).

If surface scope is in general indeed easier to access than inverse scope, what explains the converging evidence for the inverse scope preference of *every*- and *all*-negation? We argue that world expectations shaping relative plausibility help resolve this puzzle. Section 2.1 surveys the

[1] Surface scope has been called *isomorphic* (e.g., Musolino, 1999), *direct* (e.g., Ruys and Winter, 2011), *NEG-V* (e.g., Neukom-Hermann, 2016), or *high/wide scope of the first operator* (e.g., Szabolcsi, 2011); the corresponding terms for inverse scope are *nonisomorphic*, *indirect*, *NEG-Q*, or *narrow scope of the first operator*.

empirical landscape of interpretation variation, Section 2.2 develops this argument in the context of *every-* and *all*-negation, and Section 2.3 clarifies the notion of plausibility as a disambiguating factor.

2.1 Interpretation shifts by context

A striking characteristic of past findings is that identical constructions can be interpreted differently in different contexts. Indeed, changes to the local linguistic context (i.e., the sentence containing the quantifier-negation clause) can flip interpretation patterns entirely. For example, Carden (1973) found that for the sentence in (4), 92.5% of participants said that only the inverse scope interpretation was possible and 7.5% said that both surface and inverse were possible but they favored inverse scope. In other words, 100% preferred inverse scope for the quantifier-negation clause in (4). In contrast, for (5), 100% of participants said that only the surface scope interpretation was possible.

(4) All the boys didn't arrive, did they? (100% inverse scope preference; Carden, 1973)

(5) All the boys didn't leave until midnight. (100% surface scope preference; Carden, 1973)

Beyond its role in adult interpretation preferences, children's interpretation preferences are also influenced by context (Musolino, 1999; Gualmini et al., 2008; Viau et al., 2010). These studies demonstrate, in particular, that there are multiple ways to change the context to facilitate inverse scope preference. For example, in a context like (6a), children appear to have difficulty accessing the inverse scope interpretation for the utterance in (6): in a truth-value judgment task, typically less than 10% of judgments by 4-6 year-old participants endorse this utterance as a description of an inverse-verifying scenario (e.g., a scenario where two out of three horses jump over the fence; Musolino, 1999). However, in the contexts described in (6b), children increase their endorsement of the utterance in (6) to 50-60% of the time (Musolino and Lidz, 2006; Viau et al., 2010).

(6) Scenario: Two out of three horses jump over a fence.
Utterance: *Every horse didn't jump over the fence.*
   a. Context with lower endorsement of the utterance:
   previously showing a scenario in which all horses fail to jump over a barn first.
   b. Contexts with greater endorsement of the utterance:
      (i) additionally uttering "Every horse jumped over the log, but..."
      (ii) previously showing a scenario in which all horses first jumped over a log.

If different contexts lead to entirely different interpretations, then differences in context could help explain variation in interpretations. In particular, we identify a recurring pattern in the literature: contexts that favor inverse interpretations for *every*-negation and *all*-negation are those in which expectations shape the relative plausibility of competing scope interpretations. In particular, these kinds of expectations may facilitate inverse scope preference for children and adults because they make the inverse scope interpretation more plausible. Further, even allowing that surface scope is easier to access than inverse scope, *every*-negation and *all*-negation may often receive inverse scope in corpora and de-contextualized experiments because adult speakers know from experience that *every/all*-negation utterances are often used in just these inverse-facilitating contexts.

2.2 A context that makes *every*-negation inverse scope more plausible: High positive expectations

In the literature on scope ambiguity, one aspect of the context that consistently appears to facilitate inverse scope preference for quantifier-negation utterances with universal quantifiers is what we call a *high positive expectation*: the belief that the relevant entities have the property corresponding to the non-negated predicate. For example, for the *every*-negation utterance *Every vote doesn't count*, a high positive expectation would concern the prior probability that a vote counts: the greater the probability of success that votes count, the greater the high positive expectation in interlocutors' minds. As another example, for *Every horse didn't jump over the fence*, the corresponding high positive expectation is the belief that horses are likely to jump over the fence. For any *all*-negation or *every*-negation utterance, a strong version of the high positive expectation could be paraphrased by the non-negated quantifier-negation utterance itself (e.g., *Every vote counts* for *Every vote doesn't count*).

To return to the corpus-attested, inverse-scope-preferred example in (3), a high positive expectation is exactly what gets expressed in the preceding context as *everything has moved to California*. For the inverse-scope-preferred example in (4), which was given to adult participants in spoken linguistic interviews, the corresponding high positive expectation would be the belief that the boys are likely to have arrived.

Finally, a high positive expectation may have been made salient in those contexts where children's behavior suggested an inverse scope preference. For that key utterance *Every horse didn't jump over the fence*, the high positive expectation would be that horses tend to succeed in jumping. A context like (6a), which did not facilitate inverse scope preference, also did not obviously set up the expectation that horses succeed in jumping over the fence. In fact, it may have conveyed a low positive expectation, that horses are bad at jumping over things. On the other hand, the contexts described in (6b), which did facilitate inverse scope preference, perhaps communicated a high positive expectation by conveying that horses are good at jumping over things, or that the experimenters or characters in the story (who participants believe know more about the state of the experimental world than the participants do) expected every horse to jump over the fence. The higher the positive expectation, the more a participant would expect that *some jumped*, and the more surprising it would be to learn that *none jumped* rather than *some but not all jumped*. Accordingly, when *some jumped* is likely, the inverse scope *not all* interpretation is more plausible than the surface-scope *none* interpretation, which may in turn increase speakers' willingness to use *every*-negation in inverse-verifying contexts.

In their account of speaker behavior, Scontras and Pearl (2021) use an RSA model to articulate the cognitive inferences that yield observed experimental behavior for scopally-ambiguous utterances, as a way of accounting for truth value judgment patterns for quantifier-negation. In that model, world expectations make different interpretations more or less informative and thus more or less likely. The key hypothesis, which is integrated as an assumption of the model, is that a pragmatic listener assumes that a rational and cooperative speaker wants to maximize the probability that the listener will arrive at the intended understanding of the world state (while minimizing the cost of speaking). Utterances or interpretations that are more informative, in the sense that by using them the speaker will be more successful at guiding the listener to the intended world state, are reasoned to be more likely.

Given that Scontras and Pearl (2021) focused on truth-value judgments, their model is not set up to directly test whether high positive expectations make the inverse scope interpretation more likely according to listeners. Here, we address this question with our own model that builds on Scontras and Pearl's. The original model shows that when interlocutors assign a high prior probability to

the *all* world state (i.e., have a high positive expectation), the remaining *not all* states become *a priori* unlikely. As a result, the potentially ambiguous *every*-negation utterance becomes highly informative for conveying that one of these unlikely states is true, increasing speakers' willingness to use *every*-negation in inverse-verifying contexts. There are other mechanisms which can drive RSA model predictions, but this is one mechanism that Scontras and Pearl identified as particularly impactful when modeling truth value judgments.

Overall, prior research suggests that we should expect variation in scope interpretation preferences, and, while many factors matter for interpretation preferences, one factor that may facilitate the preference for inverse scope interpretations of *every*-negation and *all*-negation is a high positive expectation in the preceding context, which makes the inverse scope interpretation more plausible. Moreover, one computational hypothesis for how high positive expectations affect behavior is articulated in Scontras and Pearl's (2021) RSA model of truth value judgments for *every*-negation utterances.

2.3 Hypothesizing how world expectations affect interpretation plausibility

Taking stock, our reading of the literature suggests that the relative plausibility of competing interpretations might help account for interpretation preferences. Moreover, as Scontras and Pearl's RSA model for truth-value judgments makes concrete, world knowledge might influence relative plausibility of an interpretation in the following general way: interpretations are preferred when they are more likely to be true. This preference for true interpretations is based on the conversational goal, shared between speaker and listener, for the listener to arrive at the speaker's intended interpretation of an utterance, plus the shared conversational assumption that speakers tend to say things that are true (e.g., in accordance with the Gricean maxim of quality, to be truthful; Grice, 1975). The listener then weighs the relative probability of one interpretation over the other, according to the relative probabilities that these interpretations are true.

We note that this reasoning about plausibility concerns likely states of the world, not about specific utterance alternatives. While reasoning about what else the speaker might have said may also play a role in interpretation preferences, we focus here on how world knowledge, not utterance alternatives, shapes competing interpretations.

While we do not expect the claim that plausibility matters for interpretation to be a controversial one, our primary contribution is to make this claim concrete via a formal proposal (as implemented via our computational cognitive model). Context can seem difficult to pin down as a factor, as it encompasses any number of potentially-relevant linguistic and extralinguistic factors. Further, in the few cases (to our knowledge) when previous studies on universally-quantified quantifier-negation mention world-knowledge-in-context as a source of disambiguating information, why and how context matters are often left unexplained. For example, Neukom-Hermann (2016) suggests that the world knowledge that *Sainsbury's is a supermarket* matters for the corpus-attested case of *all*-negation in (7).

(7) Many of you may have noticed that Good Housekeeping is now on sale at the checkout in Sainsbury's, which has gone down brilliantly with shoppers, as I discovered when I visited my local London branch. I can't think why **all supermarkets don't put GH at the checkout**.

Although Neukom-Hermann does not further characterize how context might disambiguate this example, the case is in line with the broader hypothesis argued here for the role of world knowledge. Listeners can reason from *Sainsbury's is a supermarket* and *Good Housekeeping is now on sale at the checkout in Sainsbury's*, to *there exists at least one supermarket (Sainsbury's) which puts GH (Good*

Housekeeping) at the checkout. In other words, a positive expectation is set up as belief in the *some* world state. This belief means that *it is false that no supermarket puts GH at the checkout*. In other words, this belief suggests that the surface scope interpretation is false for *all supermarkets don't put GH at the checkout*. In combination with the pragmatic reasoning that cooperative speakers usually say things that are true, the listener reasons that the inverse scope interpretation, which is the only remaining interpretation which could be true, is indeed true: believing in the inverse scope interpretation is both consistent with (i.e., not ruled out by) the listener's knowledge, and it would allow the listener to interpret the speaker's meaning as true of the world.

Narrowing down to the case study of universally-quantified quantifier-negation, our reading of the literature suggests that the type of expectations that make inverse scope interpretations more plausible are high positive expectations. For a quantifier-negation sentence, the greater the expected success rate of the non-negated predicate as it applies to each entity under discussion, the higher the positive expectation. The strongest version of a high positive expectation is that worlds consistent with the truth of the non-negated expression are `true`—a belief in the truth of the *all* world state—while a positive expectation in general is a belief in the *some* (*at least one*) world state. Conversely, we would predict that a salient belief in the *none* world state would lead to a surface scope preference, all else being equal.

An open question is whether high positive expectations come into play in naturalistic speech. How often do speakers in fact say utterances such as *Every vote doesn't count*, intending the inverse scope interpretation *Not all the votes counted*, in contexts that set up the expectation that it is likely that votes count? Although quantifier-negation utterances with universal quantifiers have been studied in experiments, these studies investigated a limited number of potentially-ambiguous utterances, which may differ in many ways from quantifier-negation utterances in everyday speech. It is important to bridge the understanding of scope ambiguity that has been gained in the lab with an understanding of how ambiguity is used everyday. We know of only a few corpus studies of quantifier-negation, and one has a small sample size (Musolino et al., 2000) while the other is based primarily on written language (Neukom-Hermann, 2016) and relied on the researcher to determine the intended scope interpretation. Exactly because there is so much potential variation in preferred interpretations (what is the margin of error on a single interpretation of a single utterance?), here we use crowd-sourced interpretations of corpus-mined *every*-negation utterances to investigate the role of context for disambiguation.

The next section formally specifies our hypothesis about how world expectations affect interpretation plausibility. We build on the proposal developed by Scontras and Pearl (2021) to predict *every*-negation utterance interpretations.

3 Modeling scope interpretations

To make our assumptions explicit about how listeners resolve scope ambiguity given context, we used a computational model in the RSA framework (Frank and Goodman, 2012; Goodman and Frank, 2016). In this framework, ambiguity resolution arises from rational and domain-general inferences that listeners regularly perform as they understand language. RSA models have been shown to capture various aspects of language use (for a recent overview, see Degen, 2022). We adopt this framework as a tool for concretely implementing our hypothesis. We do not claim that ours is the only possible model of scope ambiguity resolution: alternative formalizations of how listeners integrate world knowledge and interpretive preferences could in principle yield similar predictions, though the extent to which they do so remains an open question.

For our purposes, an RSA model allows us to specify a set of assumptions about how listeners

integrate their grammatical knowledge of potential ambiguity (their knowledge of the two potential scope interpretations and a truth-functional semantics) with their goals and beliefs as social agents using language to communicate, including both world knowledge and general principles of conversation (e.g., interlocutors believe that speakers usually say things that are true and informative). In what follows, we describe our model and show how it offers an explanation for the role of plausibility, in particular high positive expectations, for *every*-negation disambiguation.

Our model adapts and extends the RSA model developed by Scontras and Pearl (2021) to account for child vs. adult behavior in past experimental work on *every*-negation. Where Scontras and Pearl focus on truth value judgments, here we model interpretation preferences directly: hearing an *every*-negation utterance, what is the probability that a listener would arrive at an inverse interpretation? The RSA model implementation reflects the behavioral task being modeled. Scontras and Pearl (2021) model truth value judgments, which capture speakers' choices about whether an utterance is an appropriate description of a given world state. In contrast, we model interpretation preferences, which capture a listener's beliefs about the world given an utterance. We therefore treat these as two perspectives within the same RSA framework, rather than as distinct models of separate phenomena. Within our model, we vary the extent to which the model assumes a high positive expectation, thereby manipulating the plausibility of the inverse interpretation.

We first specify a context where a quantity is under discussion and enumerate the possible states of the world to be described. Our communication scenario features three marbles, each one blue or red; the possible world states w to be described are defined in terms of the number of marbles that are red: $w \in W = \{0, 1, 2, 3\}$ (see Figure 1). In this scenario, a speaker tries to communicate the number of red marbles to a listener. A key component of the model is the utterance space, which captures a pragmatic factor beyond world knowledge—namely, the set of available utterance alternatives that can influence interpretation preferences. In interpreting an *every*-negation utterance, our modeled listener considers what else the speaker could have said. Following Scontras and Pearl (2021), we specify this utterance space as consisting of the potentially ambiguous *every*-negation utterance (*Every marble isn't red*) and a null utterance: U ={*every*-negation, null}.

(a) w = 0 (b) w = 1 (c) w = 2 (d) w = 3

Figure 1: Possible world states.

When interpreted with surface scope, modeled speakers and listeners understand that *Every marble isn't red* means *none are red*; when interpreted with inverse scope, they understand it means *not all are red*. For the model, this shared knowledge is reflected in the truth-functional semantics for the utterances in (8), which determines which states are `true` for a given interpretation. The semantics offers a mapping parameterized by the scope interpretation $i \in I = \{surface, inverse\}$ from world states $w \in W$ to truth values $Bool = \{\texttt{true}, \texttt{false}\}$. So, *every*-negation maps world 0 to `true` under surface scope and worlds 0, 1, and 2 (i.e., w≠3) to `true` under inverse scope. The *null* utterance does not rule out any world states, mapping all of them to `true`.

(8) Utterance semantics $[\![u]\!]^i$:

a. $[\![\textit{every}\text{-negation}]\!]^{surface} = \lambda$w. w = 0 (i.e., 'none')

b. $[\![\textit{every}\text{-negation}]\!]^{inverse} = \lambda$w. w ≠ 3 (i.e., 'not all')

c. $[\![null]\!] = \lambda$w. `true`

Given the above-specified model universe, the RSA model specifies how a listener interprets an utterance by reasoning about the speaker who generated it (and a speaker chooses an utterance by reasoning about how a listener would interpret it). Specifically, we describe how a pragmatic listener L_1 reasons about the speaker S_1 who generated the utterance, considering that S_1 was reasoning about an imagined literal listener L_0 when generating that utterance.

The hypothetical literal listener L_0 hears an utterance u and interprets it relative to its intended interpretation i; L_0 reasons that the state of the world w is any of the world states that are true, given the semantics $[\![u]\!]^i$ from (8). The model implements this reasoning as a filter on the possible world states $\delta_{[\![u]\!]^i(w)}$, which returns 1 when $[\![u]\!]^i(w)$ is `true` and 0 otherwise. L_0 then weights the true world states equally, returning a uniform probability distribution over those states w compatible with the semantics. L_0 arrives at this uniform distribution by multiplying $\delta_{[\![u]\!]^i(w)}$ (i.e., 1 or 0) by the prior probability $P_0(w)$; $P_0(w)$ represents a uniform probability distribution—the hypothesized literal listener does not have informative prior beliefs, treating all world states as equally likely.

$$P_{L_0}(w|u,i) \propto \delta_{[\![u]\!]^i(w)} \cdot P_0(w) \tag{15}$$

This assumption reflects the standard division of labor in RSA models between a literal listener and higher-level pragmatic agents. The literal listener L_0 is intended to capture only truth-conditional interpretation, abstracting away from contextual expectations and prior beliefs about the world state. These expectations instead enter at the level of the pragmatic listener L_1, where they can interact with reasoning about speaker behavior and informativity. While it would in principle be possible to encode informative priors already at L_0, doing so would blur this distinction and make it more difficult to isolate the contribution of world knowledge to pragmatic inference. In our implementation, we therefore follow prior RSA work in assuming a uniform prior at L_0, allowing prior expectations to influence interpretation preferences specifically through pragmatic reasoning.

The speaker's conversational goal in this model is to guide L_0 to the intended world state. In this setting, the goal amounts to conveying exactly how many of the three marbles are red. The speaker S_1 selects u, knowing the particular intended world w and scope interpretation i as in (16). This calculation is based on the perceived utility of u, which depends in part on the probability of u and i communicating the intended world state w to L_0: $P_{L_0}(w|u,i)$. The other component of an utterance's utility is its negative cost, $c(u)$. Broadly, utterance cost can reflect different reasons for why utterance use is difficult or effortful: for example, an utterance can be costlier than another if it is longer or less frequent. The speaker's decision process is mediated by a softmax function and free parameter α, which controls how the speaker perceives the relative contrasts between potential utilities; contrasts can be sharpened ($\alpha > 1$), smoothed away ($\alpha < 1$), or perceived as is ($\alpha = 1$).

$$P_{S_1}(u|w,i) \propto exp(\alpha \cdot \log(P_{L_0}(w|u,i)) - c(u)) \tag{16}$$

Hearing *every*-negation, a pragmatic listener L_1 reasons jointly about the true world state w and scope interpretation i that would have been most likely to lead S_1 to produce the observed utterance. L_1 considers both the prior probabilities of w and i as well as the speaker's decision process $P_{S_1}(u|w,i)$, as shown in (17). At this level, the listener's prior over world states $P(w)$ is informative, capturing expectations about which states are more or less likely in the context.

$$P_{L_1}(w,i|u) \propto P(w) \cdot P(i) \cdot P_{S_1}(u|w,i) \tag{17}$$

Note that we can specify additional layers of inference above the pragmatic listener. For example, the next layer would be a speaker S_2 as shown in (18), who observes the state of the world and chooses an utterance to convey that state of the world to L_1, marginalizing over other variables. Scontras and Pearl (2021) use S_2 along these lines to model truth-value judgments: given some state of the world (e.g., two out of three marbles are red), what is the probability of endorsing the *every*-not utterance as a description of that state?

$$P_{S_2}(u|w) \propto exp(log \sum_i P_{L_1}(w, i|u)) \tag{18}$$

Given that our focus is on interpretation preferences, we focus here on analyzing L_1 behavior, specifically the marginal posterior distribution on interpretations upon hearing the *every*-not utterance in context.

3.1 Initial parameter setting

To generate predictions from our model, we must fix the free parameters, which determine (i) the decisiveness α, (ii) the scope prior $P(i)$ (i.e., listeners' beliefs about the general probability of surface vs. inverse scope), (iii) utterance cost $c(u)$, and (iv) the world prior $P(w)$ (i.e., listeners' beliefs about the general probability of the possible world states). To implement minimal assumptions, we keep $\alpha = 1$ (that is, no sharpening or smoothing of utilities) and the prior uniform over scope interpretation such that $P(\textit{surface}) = P(\textit{inverse}) = 0.5$ (that is, neither scope interpretation is preferred a priori). With respect to utterance cost, it seems more costly to say something than to say nothing, so we set costs such that $c(\textit{every}\text{-negation}) = 1$ and $c(\text{null}) = 0$.

This leaves the world prior $P(w)$, which can implement a high positive expectation and influence interpretation plausibility. To test whether high positive expectations increase the model-predicted probability of inverse interpretations, we vary the world prior $P(w)$ (i.e., the extent to which the model assumes that marbles are red) and see the resulting predicted interpretation preference. In particular, we specify the world prior such that individual marbles have a probability p_r of being red, and each world state contains three such marbles. So, the underlying distribution for $P(w)$ is a binomial distribution with three trials, each with success probability p_r, as in (19).

$$P(w = k) = \binom{3}{k} \cdot p_r^k (1 - p_r)^{3-k} \tag{19}$$

3.2 Results

We consider the model's prediction for pragmatic listener L_1's marginal distribution over scope interpretations for the *every*-negation utterance. Figure 2 shows that the model indeed predicts that listeners should be more likely to arrive at the inverse scope interpretation of *every*-negation as their prior beliefs favor marbles being red: the higher the prior probability that a marble is red, the higher the pragmatic listener's resulting preference for the inverse scope interpretation.

3.3 Discussion

As a proof of concept, the model indeed predicts that the inverse scope interpretation of *every*-negation becomes more likely as beliefs favor high positive expectations. The formal articulation of the model also allows us to better understand why this prediction is made: it rests on the listener's reasoning that the utterance is true, and the probability that the utterance is true is higher under the inverse scope interpretation rather than the surface scope one. More specifically, there are more

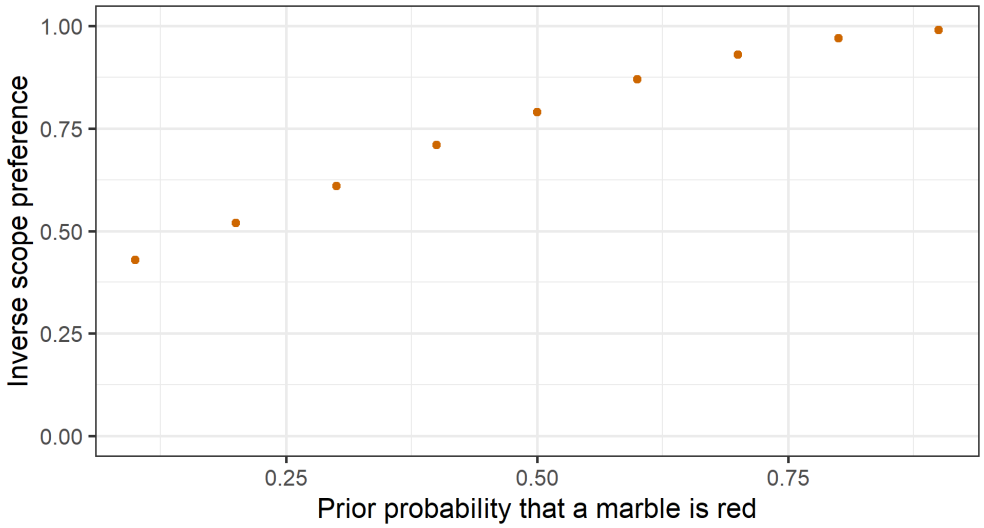


Figure 2: Predicted inverse scope preference for *every*-negation given the model's prior belief p_r that a model is red. As the probability that each marble is red rises, the extent to which there is a high positive expectation rises, and the predicted inverse scope preference also rises.

ways for inverse scope *not all* to be true (w could be 0, 1, or 2) than for surface scope *none* to be true (w must be 0). As the prior probability of a marble being red increases, the probability of world states 1 and 2 increases relative to the probability of world state 0, and so the probability placed by the pragmatic listener on the inverse scope scope interpretation correspondingly increases. In intuitive terms, the more that listeners hold a high positive expectation for *every*-negation and therefore believe there is a high probability that *some or all* is true, the more they reason that the speaker cannot have meant *none* and, therefore, meant *not all*. In other words, listeners arrive at the more plausible interpretation given their contextual expectations.

Note that this reasoning underlying the model predictions for L_1 listener behavior is different from the reasoning that Scontras and Pearl (2021) describe as underlying model predictions for S_2 speaker behavior (such as we would see in a truth value judgment task). With truth value judgments, the modeled speaker's goal is to say something as useful as possible (modeling a participant's decision to endorse or not endorse an *every*-negation utterance as a description of a scenario in which its inverse scope interpretation is true). This usefulness for the S_2 speaker is defined by informativity and cost: in particular, without varying cost, an utterance is more informative the more that the pragmatic listener's (L_1's) posterior distribution over interpretations differs from the prior distribution, and in such a way that the pragmatic listener correctly arrives at the speaker's intended interpretation. In other words, learning that a strong prior belief is false is very informative. And, since prior beliefs shift at the level of the pragmatic listener L_1, they lead to differential utility for S_2 who reasons about L_1. Thus, differential utility for S_2, operationalized via informativity, determines utterance endorsement for truth value judgments.

In contrast, with interpretation preferences, the modeled listener L_1 has the goal of reasoning about the intended interpretation of a speaker S_1, who reasons only about L_0. Prior beliefs do not shift at the level of L_0 in our model (L_0 has a flat prior on world states), so they cannot lead to differential utility for S_1. Thus speaker informativity is not affected by shifting prior beliefs when we only consider L_1 behavior; rather, it is the pressure on L_1 to reason about the ways that an interpretation can be true that is affected by shifting prior beliefs about the world.

4 Interpretation variation and plausibility in a corpus

Having specified our hypothesis for the role of plausibility in interpretation preferences, we turn to evaluating the predictions of this hypothesis. We do so by investigating *every*-negation utterances in naturalistic contexts, beginning by creating a corpus of naturalistic uses drawn from conversation transcripts (Section 4.1). We then ask naive participants to indicate their scope interpretations for the *every*-negation uses occurring in their immediate contexts (Section 4.2). Finally, to test our hypothesis that more plausible interpretations are preferred, we explore the extent to which an individual use of *every*-negation from the corpus is more likely to be interpreted with inverse scope in a context that expresses a high positive expectation (Section 4.3).

4.1 Corpus search for *every*-negation utterances

We extracted the *every*-negation occurrences in the speech section of the Corpus of Contemporary American English (COCA; Davies, 2015), defining these occurrences as those where quantified subjects precede and c-command sentential negation (with *not* or contracted *n't*). The spoken section of COCA is made up of transcripts of spoken conversations from American radio and TV programs; the license we used gave us access to ≈9 million clauses, or ≈95 million words, from 1990 to 2012.

To develop the automated search, we randomly selected a year of COCA transcripts and manually searched it for uses of *every*-negation. We then wrote a search that returned each of the occurrences in this development set. We applied this search to the rest of the COCA speech section, hand-checking the results to ascertain true hits and filter out false positives. In total, we identified 390 instances among the ≈9 million clauses searched, suggesting that *every*-negation uses are highly infrequent but do in fact occur in everyday English conversation.

4.2 Exploring naturalistic variation

We asked whether *every*-negation as attested in everyday conversation is indeed ambiguous, and what interpretation is preferred. To answer these questions, we annotated the *every*-negation corpus with crowd-sourced scope interpretations.

4.2.1 Corpus annotation

We annotated the corpus of 390 *every*-negation items with each item's preferred interpretation. Following Degen (2015), we gathered interpretations by asking participants to judge utterances in their immediate linguistic context. We measured interpretations on a sliding scale using a version of the paraphrase-endorsement methodology used by Scontras and Goodman (2017), streamlining the task because our experiment had many more trials than the original Scontras and Goodman (2017) study.

Participants. We recruited 390 participants with U.S. IP addresses and at least 95% approval ratings for at least 1,000 tasks through Amazon.com's Mechanical Turk (MTurk) crowd-sourcing service. Each participant received $2.00.

Stimuli. An example trial is shown in Figure 3. For each of the 390 *every*-negation uses in our corpus, we created an excerpt consisting of the three preceding sentences (or lines if punctuation was missing), the bolded potentially-ambiguous clause, and one following sentence (or line). For example, in Figure 3, the potentially-ambiguous clause is *Everyone does not need to establish credit*

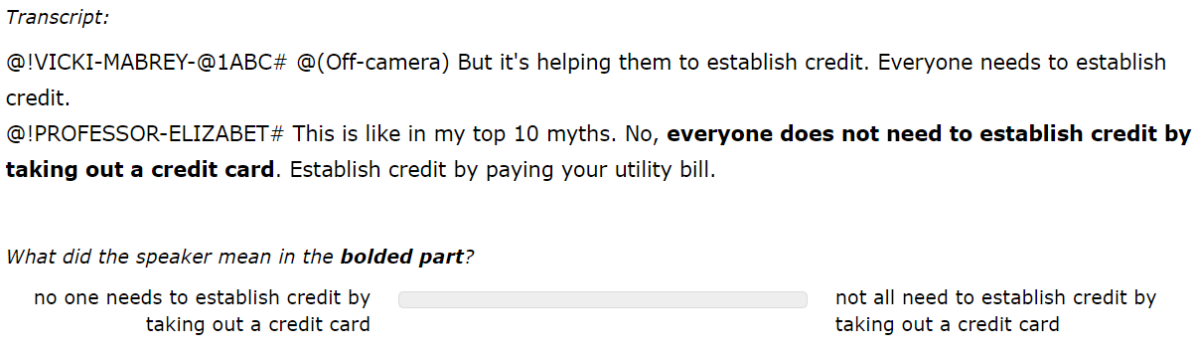


Figure 3: Sample paraphrase-endorsement trial from the corpus annotation of *every*-negation utterances.

by taking out a credit card, the preceding context is *But it's helping them ...*, and the following context is *Establish credit by ....*

For each item, we created paraphrases of the surface and inverse scope interpretations.[2] Given that the ambiguous clauses took the form *quantified noun phrase–verb–negation–remainder*, surface scope paraphrases took the form *none/no one/nobody/nothing–verb–remainder* and inverse scope paraphrases took the form *not all/not all things are–remainder*. In the example in Figure 3, the original utterance's *remainder* was *need to establish credit...*, and so the paraphrase of the surface scope interpretation was *no one needs to establish credit ...* and the inverse scope one is *not all need to establish credit ....*

Design. The initial instructions asked participants to "choose the best paraphrase for the bolded part" for fifteen randomly-selected items; on each trial, participants were again asked "What did the speaker mean in the bolded part?" (see Figure 3). Beneath the conversation excerpt, participants rated the best paraphrase as a judgment on a sliding scale between the surface and inverse scope interpretations. The two scope interpretations were randomly assigned for each item in left-right or right-left order.

Controls. To check that participants were reading and understanding the contexts of the items—and also as a way to demonstrate that context is useful for the task—two control trials were constructed to imitate the items from the corpus. The controls appeared in random order as the first two trials for each participant. These control trials contained clearly disambiguating information about the intended scope interpretation in the surrounding context. The disambiguating information always appeared as a restatement of the speaker's meaning.

The surface scope-disambiguating control item is in (9), and the inverse scope-disambiguating control item is in (10). For clarity, the disambiguating information is italicized, though it was not italicized in the experiment. Participants were considered to pass the surface control by placing the slider closer to the *none* paraphrase than to the *not all* paraphrase; they passed the inverse control by placing the slider closer to the *not all* paraphrase than to the *nobody* paraphrase.

(9) TONHAUSER: The ten board members voted last night. I was really surprised—I thought at least some of them would like Proposition 23. But *all ten of them voted against it.* Basically, **every board member didn't like Proposition 23**. *Not even a single one of them liked it.*

[2]The form of these paraphrases was validated in a separate experiment, described in Section 5.2.1 below.

(10) SIDNER: Look, we completely fixed the issue. Indicators have improved across the board. Everybody's happy.
GROSZ: (VOICEOVER) No, **everybody isn't happy**. *Some are happy but others are deeply dissatisfied with what they call a 'band aid solution.'*

The rate of passing both controls was 53%. This relatively low pass rate may have been due to low English reading proficiency, low attention and motivation, or high task difficulty. Though we restricted MTurk participation to US IP addresses and to those MTurk workers who have completed at least 1,000 tasks in the past, and we also only analyzed data from self-reported native English speakers, some participants may not have fluently read English well enough, or they may have lacked motivation or engagement to read the items in detail. Participants in an online study, or on the MTurk platform in particular, may be disengaged with the experiment. A third factor is task difficulty: the paraphrase endorsement task is a kind of complex reading comprehension and logical inference task, because these sentences have multiple logical operators.

With the addition of the two controls, participants completed a total of 17 trials. We restricted analysis to those participants who passed both controls and indicated English as their only native language. Out of the 390 participants, we assessed data from 208 (35% female; mean age: 41).

4.2.2 Results

Each item was judged by at least 2 and at most 14 different participants, with an average between 8 and 9 ratings per item. Although the surface scope paraphrases randomly appeared on the left or right of the sliders, we transformed and report responses on sliders as though the surface scope paraphrases always appeared on the left. As a result, the final response measure for each trial varies from 0 (maximum endorsement of the surface scope interpretation) to 1 (maximum endorsement of the inverse scope interpretation).

As shown in Figure 4, we found both a general preference for inverse scope interpretations and a high degree of interpretation variation for the corpus *every*-negation utterances. The left panel of Figure 4 shows judgment-by-judgment interpretations, and suggests that many of these utterances in context elicit strong intuitions such that they are indeed unambiguous in context: 29% of individual scores were below 0.25 (indicating a strongly surface scope interpretation) while 53% of individual scores were above 0.75 (indicating a strongly inverse scope interpretation). The right panel of Figure 4 shows the mean interpretations per item, and suggests that for some of our items, these strong intuitions are reliable across different participants' judgments: 12% of mean scores were below 0.25, and 38% of mean scores were above 0.75.

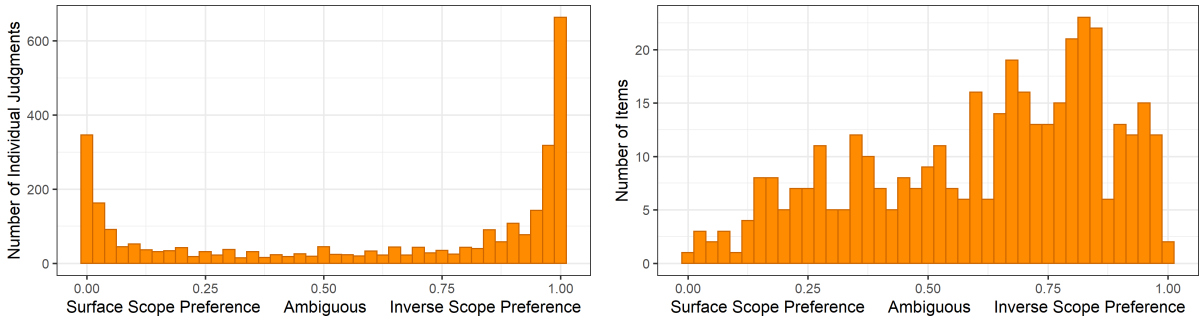


Figure 4: Individual scope interpretations (left panel) and mean interpretations per item (right panel) from the *every*-negation corpus analysis.

Figure 5 shows examples of four attested interpretation patterns: a strong surface scope preference for the item in (11) (top slider; mean response ≈ 0), a strong inverse scope preference for the item in (12) (second slider; mean response ≈ 1), and the two forms of true ambiguity (mean response ≈ 0.5). In (13) (third slider in Figure 5), we see high cross-rater disagreement, and in (14) (fourth slider in Figure 5) we see high cross-rater agreement. This last interpretation pattern is actually quite rare; in general, participants rarely placed the slider at the midway point between the two interpretation paraphrases, as is evident in the left panel of Figure 4.

(11) (BEGIN VIDEO CLIP, SEPTEMBER 19, 2001) HOWARD LUTNICK, CEO, CANTOR FITZGERALD: Every person who came to work for me in New York, **everyone that was in the office isn't there anymore**, every single one who was there isn't there anymore. You can't find them.

a. No one (that was in the office) is there anymore. *(every > n't)*
b. Not all (that were in the office) are there anymore. *(n't > every)*

(12) HOWARD KURTZ: At the risk of suggesting that this is not, perhaps, one of the great technological breakthroughs of the late 20th century, like, say, the microwave oven, the level of hype here has been incredible. I mean buying up 1.5 million copies of the London Sunday Times and giving them out for free? The press has- there's this fascination with high-tech computer subjects. We sometimes forget that **everybody in the world is not on-line**, is not going to go out and buy Windows. @ @ @ @ @ @ @ @ @, what does this tell us about the journalistic mind set, this hype?

a. Nobody (in the world) is on-line. *(every > n't)*
b. Not all (in the world) are on-line. *(n't > every)*

(13) Instead, he badmouths people, insults people, and has a crass attitude toward anyone who has got problems, or is weaker than he is as a governor and a wrestler. And I do @ @ @ @ @ @ @ @ @ money from it. I think it's unethical.
@!MAN: **Everything I've heard him say has not been ... good**, you know, hasn't been right.
@!MAN: I personally don't think he's taken much time to be governor.

a. Nothing (that I've heard him say) has been good. *(every > n't)*
b. Not all things (that I've heard him say) have been good. *(n't > every)*

(14) Just one week ago, Education Secretary Richard Reilly reported that 90 percent of America's schools like Jonesboro were free from violence. Now Jonesboro has become the sixth time students have fired on fellow students and teachers in the last two and a half years. And Congress is already talking about new laws to prevent another one.
@(BEGIN-VIDEO-CLIP)
@!SEN-DICK-DURBIN-@: There is no reason why[3] **every child in America shouldn't be protected at least in some small way**, by assuming that every owner of a gun has to own it responsibly, keep it in a safe manner, keep it in a way where it can not be accessed by children.
@(END-VIDEO-CLIP)
@!PRESS: Is it that simple?

a. None (in America) should be protected at least in some small way. *(every > n't)*
b. Not all (in America) should be protected at least in some small way. *(n't >every)*

[3]It's worth noting that this preceding linguistic structure "There is no reason why" may have made interpreting this item more difficult or confusing to the participants.

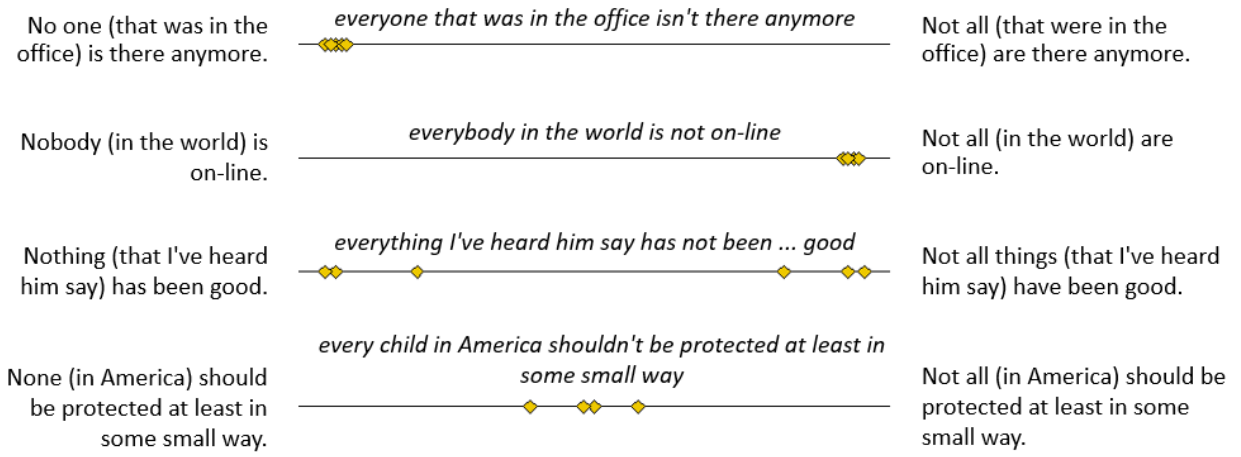


Figure 5: Individual interpretations of (11) (top slider), (12) (second slider), (13) (third slider), and (14) (fourth slider). In this figure, the horizontal line represents the length of a slider and each yellow diamond represents an individual judgment. The responses for these four items demonstrate four types of judgment patterns: unambiguous preference for surface scope (top slider) and inverse scope (second slider), ambiguity which reflects judgment disagreement (third slider), and true ambiguity on an individual judgment basis (fourth slider).

4.2.3 Discussion

We found that naturalistic *every*-negation utterances are attested and ambiguous in context, with a general preference for the inverse scope interpretation. Specifically, different uses receive a range of average interpretations, though inverse scope interpretations dominate.

Perhaps the main takeaway from the results of our corpus annotation is the variability in interpretations we document for a single type of utterance, *every*-negation. This picture of ambiguity only emerges when we consider interpretations by many participants for many different items, which demonstrates the value of a naturalistic corpus and crowd-sourced annotations. In general, to better understand linguistic ambiguity, this corpus study shows the value of data that include multiple instances of the same type of ambiguity and multiple judgments of each instance of the ambiguity. As this case study of *every*-negation shows, an individual judgment for a single utterance may not provide enough information about how people prefer to interpret that type of utterance.

Next, we explore the extent to which our hypothesis concerning plausibility can help us make sense of some of the variability in our annotated corpus.

4.3 Plausibility in the corpus

We expect that plausibility will help account for some of the variation in interpretation preferences for *every*-negation utterances in the speech corpus. In particular, in line with our model predictions, we expect that high positive expectations make inverse scope interpretations more plausible and therefore preferred. That is, we expect to find that an item was more likely to receive an inverse scope interpretation in a context containing a high positive expectation. To test this prediction, we looked for expressions of high positive expectations in the contexts of the corpus items and measured whether these expressions predicted inverse scope preference in the crowd-sourced interpretations.

4.3.1 Coding for positive expectations in corpus contexts

One way to measure for the salience of a high positive expectation is by its overt linguistic expression in context. For the high positive expectations of *every*-negation, this overt linguistic expression can come in the form of the non-negated utterance itself, which in fact would express a strong version of the high positive expectation. For example, for *Every vote doesn't count*, a high positive expectation is the prior belief that votes *do* count. One unambiguous, strong version of this belief would be expressed by the expectation that it is highly probable that *every* vote counts. So, we would know that this expectation is salient for interlocutors if it were expressed as the non-negated counterpart *Every vote does count* in the preceding context of *Every vote doesn't count*.

As a preliminary measure, the first author hand-coded categorically for the presence/absence of this kind of overt high positive expectation expression in the preceding context of each of our items. 59/390 (15%) of the items contained such an expression (that is, the non-negated counterparts of the potentially-ambiguous utterances).

4.3.2 Crowd-sourcing annotations for positive expectations in COCA

The categorical method of identifying positive expectations in the text may fail to capture relevant expectations not overtly expressed in the preceding linguistic context. Thus, we also measured positive expectations by crowd-sourcing judgments of the success rate of the non-negated predicate. This crowd-sourced behavioral measure captures world knowledge in addition to what is specifically expressed in the preceding context. By capturing beliefs as gradient rather than categorical, the crowd-sourced measure may provide a more sensitive alternative to the hand-coded measure.

In the crowd-sourced measure, the preceding contexts of the 390 *every*-negation items (without the *every*-negation sentences themselves) were annotated with people's judgments of the extent to which the context contained a positive expectation. As before, judgments were measured on a sliding scale.

Participants. 347 participants were recruited through Prolific, who had U.S. IP addresses and indicated that they were monolingual English speakers. Each participant received $2.00.

Stimuli. Figure 6 shows example trials. Participants saw excerpts consisting of the three sentences (or lines if punctuation was missing) that preceded the *every*-negation item. Beneath, participants saw a question intended to measure how strongly they held a positive expectation given the context, in the form of *How likely is it that a random ...* combined with the non-quantified subject and non-negated predicate of the original *every*-negation item (e.g., *How likely is it that a random person is sitting home waiting for some pollster to call?*). Given that the ambiguous clauses took the form *quantified noun phrase–negation–verb–remainder*, the question had the form *How likely is it that a random–noun phrase–verb–remainder*. Participants gave a judgment on a scale between "very unlikely" (always on the left) and "very likely" (always on the right).

In some cases, there was not enough content in the *every*-negation construction to fully formulate the positive expectation question. For example, for an item of the form *Everybody's not*, the positive expectation question is *How likely is it that a random person is?*, which is either misleading or unclear. To avoid this potential confusion, questions based on *every*-negation items which contained anaphoric or elided elements were replaced wherever possible with their antecedent in the preceding context. An example is shown in Figure 6, right panel. The original utterance was *Everybody's not*, said in response to the other speaker's claim that *everybody is worried about the amount of*

this expense. For this case, we formulated the positive expectation as *a random person is worried about the amount of this expense*.

Design. The initial instructions told participants, before they accepted the task, that "You will be asked to judge statements about the world, given short written excerpts from real conversations." After participants accepted the task, more in-depth instructions then said, "You will see short excerpts of conversations from American radio and TV programs that took place between 1990 and 2012. With each conversation, you will also see a statement about the world. **Based on the conversation and your own knowledge of the world, your task is to judge how likely that statement is to be true.** Because we're peeking into the middle of real conversations, sometimes it may be hard to know exactly what the speakers were talking about, or it may be hard to see the connection between the conversation and the statement about the world. That's okay! Please indicate your best guess." There also followed several brief statements about the form of the excerpts and the fact that some conversations were about sensitive topics.

Each participant saw a total of fifteen randomly-selected items; on each trial, participants were again asked "How likely is it that *positive expectation*?" (see Figure 6).

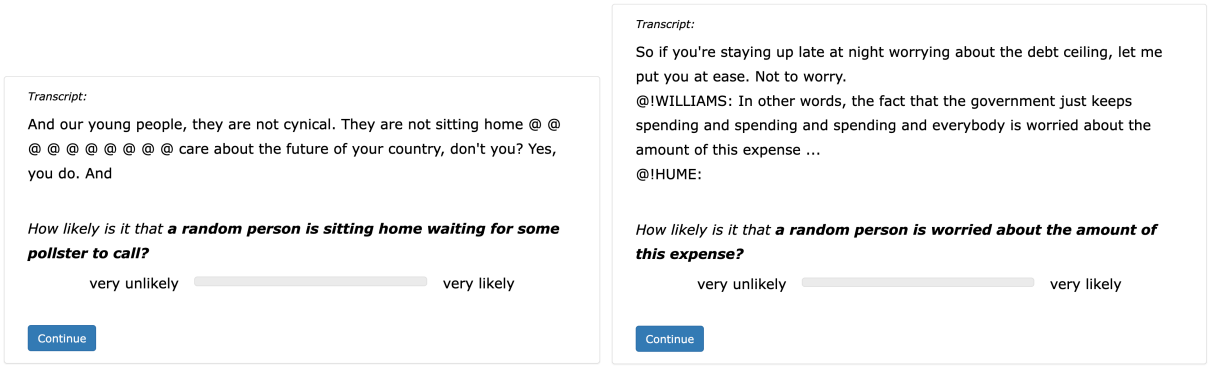


Figure 6: Sample trials from the crowd-sourced context annotation of *every*-negation utterances. The left panel shows the context for the original item "everybody is not sitting home waiting for some pollster to call". The right panel shows the context for the original item "everybody's not".

Controls. To check that participants were paying attention, reading the contexts, and understanding the task, three control trials were constructed to imitate the rest of the items. Two of the controls appeared in random order as the first two trials for each participant, and the last control item appeared as the last trial for each participant, mainly to check continued attention at the end of the task. These control trials contained clear information answering the question about the probability of a positive expectation.

The first two control items are described below in (15)—a low-probability control—and (16)—a high-probability control. For clarity, the text containing the key information (answering the question about the probability of a positive expectation) is italicized here, though it was not italicized in the experiment. Participants were considered to pass the low-probability control by placing the slider closer to the *very unlikely* side than to *very likely*; they passed the high-probability control by placing the slider closer to the *very likely* side than to the *very unlikely* side.

(15) @!TONHAUSER: The ten board members voted last night. No surprise - *every single one of them hated Proposition 23. All ten of the board members voted against it.* Basically,

How likely is it that ***a random board member liked Proposition 23****?*

(16) @!SIDNER: I'm glad to report that they completely fixed the issue at Greenwell. *Indicators have improved across the board and everybody's smiling. Everybody's happy.*

@!GROSZ: (VOICEOVER) Yep,
How likely is it that ***a random person at Greenwell is happy****?*

The third control item, which checked attention on the last trial, is described below in (17) and was a high-probability control, like (16).

(17) @!ROBERTS: You know, they said that that they completely fixed the issue at Silver Lake. *They reported improved indicators across the board and apparently everybody's smiling. I heard everybody's happy.*
@!ARIEL: (VOICEOVER): I heard the same thing,

How likely is it that ***a random person at Silver Lake is happy****?*

The rate of passing all three controls was 72%.

With the addition of the three controls, participants completed a total of 18 trials. Analysis was restricted to those participants who passed all three controls. Out of the 347 participants, data was assessed for 250 (54% female; mean age: 39.6 years).

4.3.3 Results

Hand-coded results. Of the 59 utterances that were identified via hand-coding to have high positive expectation expressions, 50/59 (85%) were on average better paraphrased by the inverse scope paraphrase than the surface scope paraphrase according to our crowd-sourced annotators.

We also looked at p(high positive expectation|inverse) vs. p(high positive expectation|surface): how often items where the inverse interpretation was strongly preferred had a high positive expectation expression compared with items where the surface interpretation was strongly preferred. We found that 22% of highly inverse-preferred items (those with responses greater than 0.75) had high positive expectation expressions, as opposed to 6% of highly surface scope-preferred items (those with responses less than 0.25). These results suggest that the hand-coded high positive expectations do tend to co-occur with an inverse scope interpretation in our sample, providing some support for our hypothesis that high positive expectations help yield inverse interpretations.

Crowd-sourced results. Each item's preceding context was judged by at least 3 and at most 16 different participants, with an average of between 9 and 10 ratings per item. The final response measure for each trial varies from 0 (maximum endorsement that a positive expectation is very unlikely) to 1 (maximum endorsement that a positive expectation is very likely). As with the hand-coded measure of positive expectations, we asked whether an item was more likely to receive an inverse scope interpretation in a context containing a higher positive expectation according to the crowd-sourced measure.

The left panel of Figure 7 shows individual judgments of positive expectations, which peak at the endpoints and midpoint of the scale (i.e., 0, 0.5, and 1), with a greater preference to place the slider below 0.5 than above it. The right panel of Figure 7 shows the mean context judgments per item, peaking slightly below the midpoint.

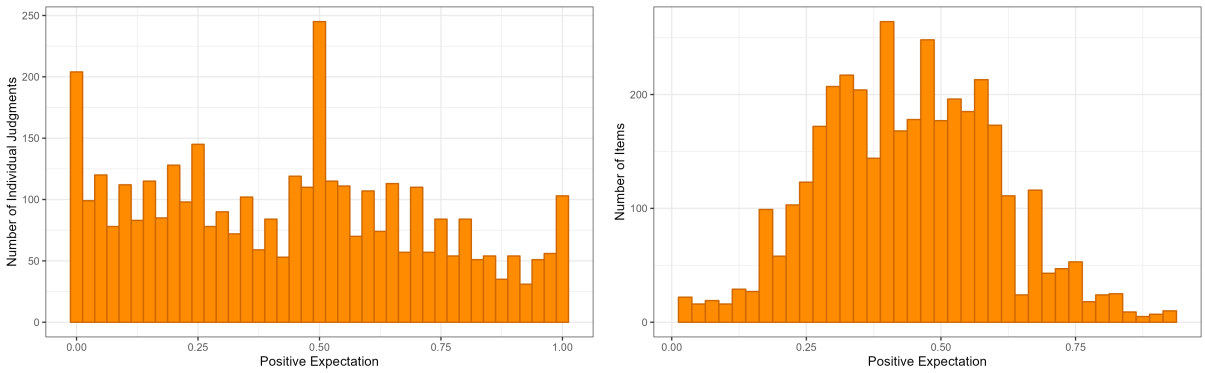


Figure 7: Individual judgments of positive expectations in context (left panel) and mean judgments per item (right panel) for the *every*-negation items from COCA.

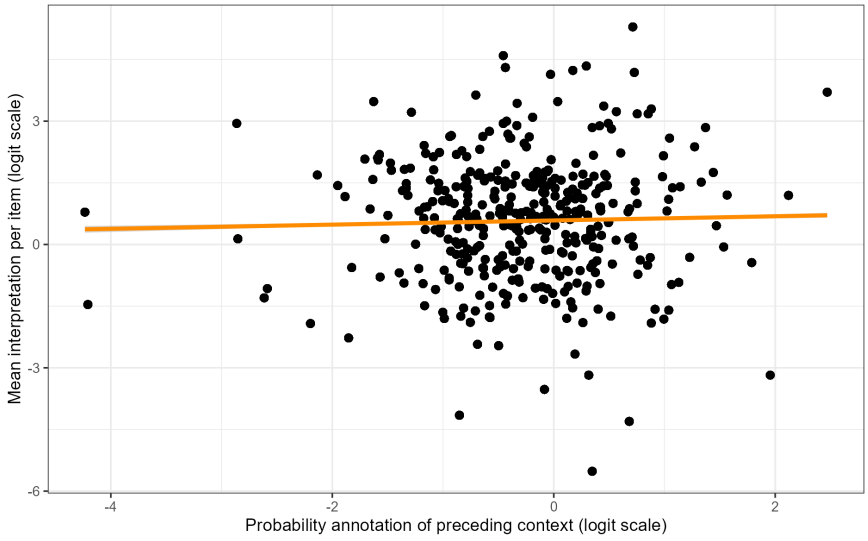


Figure 8: Positive expectation annotation of the preceding contexts for the *every*-negation items from COCA, predicting mean item inverse scope preference.

The effect of crowd-sourced positive expectation is significant and in the expected direction, though modest. Figure 8 shows mean item inverse scope preference by positive expectation annotation. To assess significance, we used a mixed effects model predicting scope interpretation by context annotation, with random intercepts for participants as the maximal random effects structure supported by the data. Model results are shown in Table 1. The intercept of 0.5887 represents the log-odds of the mean scope preference when the mean context annotation is at its reference level (log-odds = 0); using the inverse logit function to transform the log-odds back into probabilities, the reference predicted scope preference is 0.64—in other words, the item is already somewhat likely to have inverse scope. The coefficient for the context annotation predictor means that for each one-unit increase in the log-odds of the context annotation, the log-odds of the mean scope preference increases by 0.04312; applying the inverse logit (to 0.5887+0.04312=0.63182), we get 0.65, meaning that a one-unit increase in the log-odds of context annotation leads to a predicted probability of approximately 0.65—in other words, with a higher positive expectation, the item is slightly more likely to have inverse scope.

Fixed Effects (log-odds)	Estimate	Standard Error	t-value	p-value
Intercept	0.5887	0.02707	21.748	<2e-16
Crowd-sourced Positive Expectation	0.04312	0.01036	4.163	3.15e-05

Table 1: Results of a mixed effects model with mean crowd-sourced context annotation per item (log-odds; higher values indicating higher positive expectation) predicting mean scope preference per item (log-odds; higher values indicating greater inverse scope preference), with random intercepts for the participants in the context annotation experiment.

4.3.4 Discussion

We find that scope interpretation preferences of an individual use of *every*-negation from the corpus depend in part on whether its preceding context expresses a high positive expectation. In particular, high positive expectations correlate with stronger preferences for the inverse scope interpretation. These results align with our predictions from Section 3 regarding the role of positive expectations. The world expectations make the inverse scope interpretation relatively more plausible than the surface scope interpretation, such that a context with a positive expectation is one reason why a listener would prefer the inverse scope interpretation of an *every*-negation use.[4]

The correlation we observe is a modest one, which may be due to our methods of identifying a high positive expectation in context. On the one hand, both measures—categorical high positive expectation and crowd-sourced positive expectation—demonstrate the predicted positive relationship with inverse scope preference. The advantage of the categorical annotation is that it clearly marks cases of salient high positive expectation, and indeed shows the strongest correlation with inverse scope. However, it only captures cases in which a very high positive expectation is transparently encoded in the linguistic context, and therefore misses cases where world expectations are not expressed, only indirectly expressed, or expressed outside the immediate context.

In contrast, the crowd-sourced measure better captures positive expectations, as it is not limited to overt expressions or specific linguistic forms. That being said, a potential disadvantage of the measure is that the task to elicit it may have been difficult or confusing to participants. One of the peaks in the distribution of responses estimating this positive expectation is at approximately the midpoint between very unlikely and very likely, which may have reflected that in many cases participants were unsure of the probability.

Still, the advantage of the crowd-sourced annotation is that it provided a continuous measure to improve our analysis of larger-scale data. Here, it allowed us to confirm the model-predicted relationship between the strength of positive expectations and the extent of an inverse preference.

[4]We expect that multiple factors play a role for interpretations, and our argument here is that plausibility is one of those factors. However, one concern about these findings is that the observed relationship between context and scope interpretations might be reducible to the speaker's use of overt denial, rather than to plausibility. Overt denial markers were used in the constructed example in (10) and in the corpus example shown in Figure 3. We might speculate that a denial marker signals the speaker's intention to indicate a disagreement best paraphrased by the inverse scope interpretation.

To investigate the role of overt denial, we annotated each corpus item for whether the speaker used overt denial immediately preceding their use of *every*-negation. We found 44/390 cases of the denial markers "no", "however", "on the other hand", or "but". We fit a linear mixed-effects model predicting mean scope interpretations by both the presence of overt denial and the mean crowd-sourced rating of positive expectations in context, including a random intercept for participant. Both overt denial and positive expectations were significant predictors of inverse scope. In other words, crowd-sourced positive expectation ratings remained a reliable predictor of interpretations; this effect of positive expectations was only modestly reduced compared to the model without overt denial. This persistence of the effect of expectations in context demonstrates that, while overt denial contributes to interpretations, it does not subsume the role of plausibility.

5 Extending the model to different quantifiers

Our model-implemented hypothesis about scope ambiguity resolution demonstrates how high positive expectations can explain some of the observed interpretation variation for naturally-occurring *every*-negation utterances in context. Because the model's mechanism of ambiguity resolution involving world priors is intended to be general, here we assess whether our model can account for quantifier-negation interpretation preferences with other quantifiers.

We investigate the quantifiers *some* and *no*, because universal *every*, existential *some*, and negative *no* fall into three different classes (e.g., according to the classification system in Beghelli and Stowell, 1997). Intuitively, we expect these three kinds of utterances to have different preferred interpretations. For example, *some* is generally expected to scope above negation (Szabolcsi, 2004), so we expect *some*-negation to usually or always receive a surface scope interpretation (because its inverse scope interpretation involves negation scoping over *some*). The predictions for *no*-negation utterances are less clear, in part owing to the difficulty introduced by double negation.

To extend the model—and thereby explore the generalizability of our hypothesis—and generate testable predictions, we modify the model space of utterances and semantics to include *some*-negation and *no*-negation. Making minimal assumptions, we then describe the predicted interpretation preferences.

5.1 Extended model articulation

We update the set of utterances and their corresponding semantics to include *some*-negation and *no*-negation. A speaker chooses to say one of the potentially-ambiguous quantifier-negation utterances $u \in U = \{$*every*-negation, *some*-negation, *no*-negation, *null*$\}$; in other words, speakers can say *Every marble isn't red*, *Some marble isn't red*, or *No marble isn't red*, or they can say nothing at all.

Speakers and listeners have the following interpretations, as shown in the truth-functional semantics in (20):

- *Every marble isn't red* means *none are red* when interpreted with surface scope and *not all are red* when interpreted with inverse scope.
- *Some marble isn't red* means *not all are red* when interpreted with surface scope (i.e., there is some marble that is not red). It means *none are red* when interpreted with inverse scope (i.e., it is not the case that there is some red marble).
- *No marble isn't red* means *all are red* when interpreted with surface scope (i.e., for no marble is it the case that that marble is not red). It means *some are red* when interpreted with inverse scope (i.e., it is not the case that no marble is red, so at least one is red).

(20) Utterance semantics $[\![u]\!]^i$:

a. $[\![$*every*-negation$]\!]^{surface} = \lambda$w. w $= 0$ (i.e., 'none')
b. $[\![$*every*-negation$]\!]^{inverse} = \lambda$w. w $\neq 3$ (i.e., 'not all')
c. $[\![$*some*-negation$]\!]^{surface} = \lambda$w. w $\neq 3$ (i.e., 'not all')
d. $[\![$*some*-negation$]\!]^{inverse} = \lambda$w. w $= 0$ (i.e., 'none')
e. $[\![$*no*-negation$]\!]^{surface} = \lambda$w. w $= 3$ (i.e., 'all')
f. $[\![$*no*-negation$]\!]^{inverse} = \lambda$w. w > 0 (i.e., 'some')
g. $[\![$*null*$]\!] = \lambda$w. `true`

All other aspects of the model articulation remain the same.

5.1.1 Extended model parameter setting

As before, given this model articulation, we have freedom to vary the decisiveness α, the scope prior $P(i)$, the world prior $P(w)$, and the utterance costs $c(u)$. To implement minimal assumptions, we keep $\alpha = 1$ and the prior uniform over scope interpretations ($P(\textit{surface}) = P(\textit{inverse}) = 0.5$). For $P(w)$, we set the base rate of marbles being red at $p_r = 0.5$, such that a marble is equally likely to be red or not.

For utterance costs, we maintain the assumption that to say nothing costs less than to say something (cost(*null*) = 0 < cost(*every/some/no*-negation)). In addition, we set the relative costs of *every*-, *some*-, and *no*-negation to reflect their relative frequency in speech, such that less frequent utterances cost more.[5] To estimate appropriate values, we used the methodology described in Section 4.1 for mining *every*-negation from a speech corpus to also mine *some*-negation and *no*-negation utterances from COCA. We identified 2,947 occurrences for *some*-negation and 50 occurrences for *no*-negation. We set the relative costs of the utterances as inversely proportional to their relative frequency in the corpus, given the previous 390 *every*-negation instances we found: cost(*every*-negation) $= \frac{1}{\frac{390}{390+2947+50}} = 8.684615$, cost(*some*-negation) $= \frac{1}{\frac{2947}{390+2947+50}} = 1.149304$, cost(*no*-negation) $= \frac{1}{\frac{50}{390+2947+50}} = 67.741$.

5.1.2 Extended model predictions for scope interpretation preferences

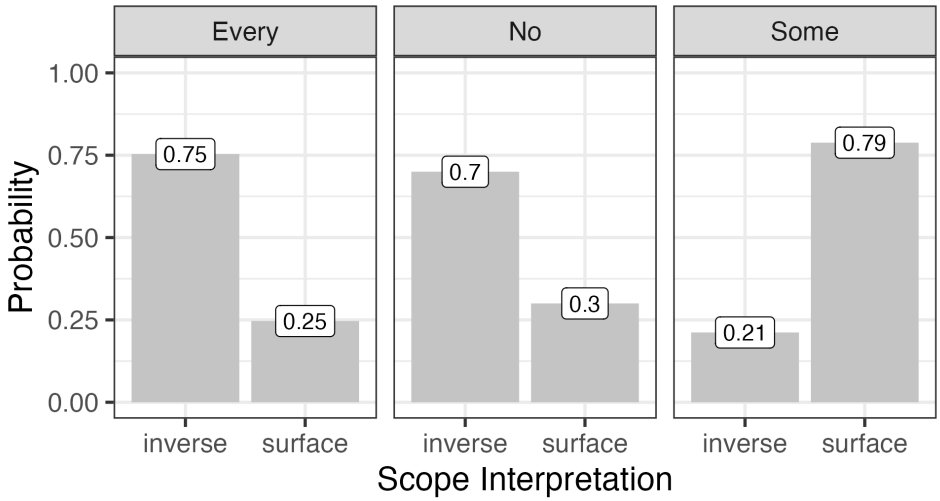


Figure 9: L_1 marginal probability distribution over scope interpretations, when we only assume relative utterance costs that reflect their relative frequencies of use in spontaneous speech (i.e., the rare *no*-negation is highly costly, *every*-negation moderately costly, and the relatively most common *some*-negation is slightly costly; to say nothing costs nothing). Otherwise, $\alpha = 1$, the prior over scope interpretations is uniform, and each marble has a 50% chance of being red $p_r = 0.5$.

[5] We use higher attestation frequency as a correlate with lower production cost because higher attestation frequency generally indicates higher degree of automatization, shorter naming latencies, higher identification accuracy, and higher predictability (for a review, see, e.g., Hilpert, 2025). On the other hand, we acknowledge that the corpus frequencies are an imperfect estimator of production cost, as they do not capture individual variation of experience with relevant words and are only as representative as the speech section of COCA can be claimed to be of naturalistic English. In addition, production cost is influenced by factors including thematic context, degree of activation in prior discourse, and phonological similarity.

Figure 9 shows the model's predicted interpretation preferences for each quantifier-negation type. Under these parameter settings implementing minimal assumptions, the model predicts that the proportion of inverse scope interpretations depends on the quantifier. The probability that *every*-negation receives an inverse scope interpretation (0.75) is greater than the probability that *no*-negation receives an inverse scope interpretation (0.7), which is greater than the probability that*some*-negation receives an inverse scope interpretation (0.21).

Specifically, the *not all* interpretation is the preferred scope interpretation for *every*-negation (inverse=0.75) and *some*-negation (surface=0.79). The reason is the same for both quantifiers, and is the same as that described in Section 3.3. Since we set the marble redness base rate to $p_r = 0.5$ (i.e., chance), the most likely world states are those where exactly one or exactly two marbles are red (as shown in Figure 10). It is more likely for *not all* to be true (w could be 0, 1, or 2, and world states 1 and 2 are relatively most likely according to our prior) than for *none* to be true (w must be 0, and 0 is relatively unlikely according to our prior). In other words, the *not all* interpretation is more plausible than the *none* interpretation, given these expectations about the world. The listener reasons that the utterance is true, and so reasons that the speaker most likely intended the meaning that is more plausible: the *not all* meaning (i.e., inverse for *every*-negation and surface for *some*-negation).

For the same reason, the *some* interpretation is preferred over the *all* scope interpretation for *no*-negation (inverse=0.7). In particular, it is more likely for *some* to be true (w could be 1, 2, or 3, and world states 1 and 2 are relatively most likely according to our prior) than for *all* to be true (w must be 3, and 3 is less likely according to our prior). The listener reasons that the speaker most likely intended the meaning that is more plausible: the *some* meaning.

Let us put these predictions again in intuitive terms, given these minimal-assumption model parameters where the most likely world states are the *some but not all* ones *a priori*. Upon hearing *every*-negation or *some*-negation, listeners will believe there is a high probability that *some but not all* is true; so, the speaker cannot have meant *none* and, therefore, meant *not all* instead (i.e., the inverse scope interpretation of *every*-negation and the surface scope interpretation of *some*-negation). Upon hearing *no*-negation, listeners will believe there is a high probability that *some but not all* is true, such that the speaker cannot have meant *all* and, therefore, meant *some* instead.

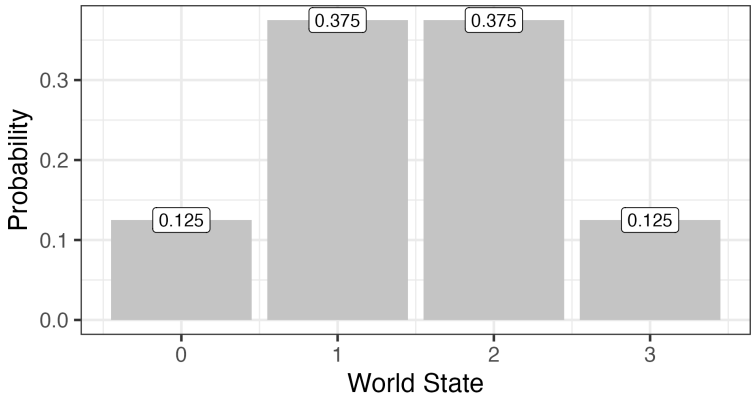
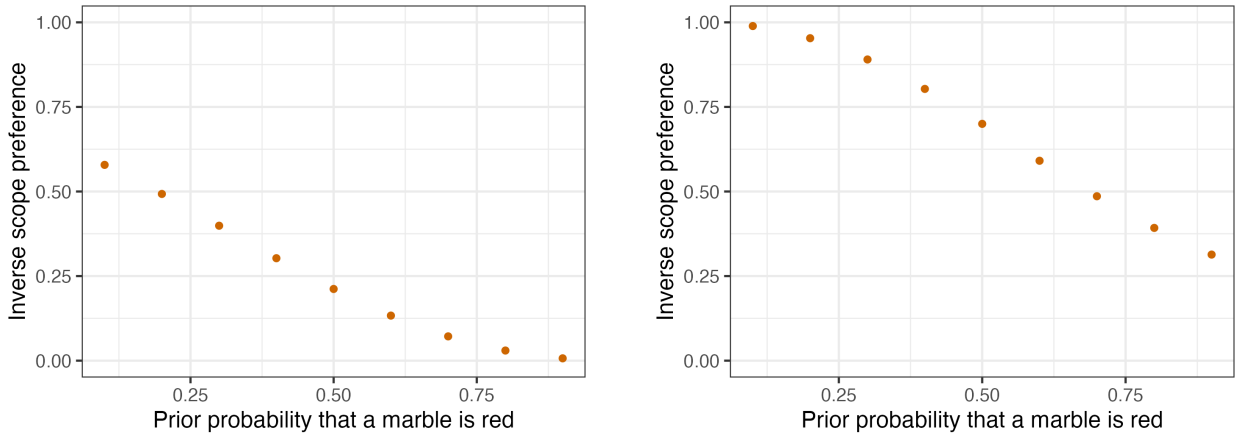


Figure 10: Prior probability distribution over world states when $p_r = 0.5$.

As a demonstration of how the model predicts that positive expectations affect plausibility (and therefore the preferred interpretation of an utterance), Figure 11 shows that listeners should be less likely to arrive at the inverse scope interpretation of *some*- and *no*-negation as their prior beliefs favor marbles being red. For *some*-negation, the higher the prior probability that a marble is red,

as mentioned above, the more that the *not all* (surface scope) rather than the *none* (inverse scope) meaning becomes relatively more plausible. For *no*-negation, the higher the prior probability that a marble is red, the more that the *all* (surface scope) rather than the *some* (inverse scope) meaning becomes more plausible, as greater probability is shifted to the *all* world state.



(a) L_1's interpretation of *some*-negation

(b) L_1's interpretation of *no*-negation

Figure 11: Predicted inverse scope preference for *some*- and *no*-negation given the model's prior belief p_r that a model is red. As the probability that each marble is red rises, the surface scope interpretation becomes relatively more plausible for both utterances.

With these predictions from the unfit version of our extended model in hand, we next see whether the predictions are borne out in human interpretation patterns. If they are, we have more general support for our model-implemented hypothesis about the role of plausibility in scope disambiguation.

5.2 Testing model predictions for *every*-, *some*-, and *no*-negation

To establish an empirical baseline for across-quantifier variation, we elicited native English speakers' average interpretation preferences for utterances with the quantifiers *every* vs. *some* vs. *no*. This baseline serves as a comparison for testing the predictions of our extended model. The stimuli were these three quantifier-negation utterances with no linguistic context, embedded in a communication scenario with two characters. In a reference picture-selection experiment, we first validated that the relevant paraphrases of each potentially-ambiguous utterance were understood to have a meaning compatible with surface- vs. inverse-verifying scenarios.

5.2.1 Paraphrase validation

Following the methodology of Scontras and Goodman (2017), we first verified unambiguous paraphrases of our potentially ambiguous utterances. We asked participants, given a paraphrase, to select the picture that the paraphrase likely described (see Figures 12 and 13).

Participants We recruited 102 participants with U.S. I.P. addresses through MTurk. Each received \$0.50. 94 participants (42% female; mean age: 37) indicated that they understood the experiment and that English was their only native language; their data were included in the analyses reported below.

Design The experiment began with a scenario intended to establish that the utterances to be interpreted were communication acts (Figure 12). A character, Mellow, is said to have a collection of marbles, three of which she places into a box. Participants were told that Mellow tells another character, Bluesy, about the box of marbles, and that their task is to help Bluesy interpret Mellow's utterance.

Participants then saw in random order three trials where they chose the scenario they thought an utterance described: one trial for the quantifier-negation utterance, one for its surface scope paraphrase, and one for its inverse scope paraphrase. The quantifiers *every*, *some*, and *no* were tested as a between-subject manipulation. On each trial, participants chose between an image consistent with the surface scope interpretation of the quantifier-negation utterance and an image consistent with the inverse scope interpretation (e.g., a participant in the *every*-negation condition chose between not-all-red-marbles and no-red-marbles, as in Figure 13); image position (left vs. right) was randomized on each trial.

The surface/inverse scope paraphrases appear in (21) for *every*, (22) for *some*, and (23) for *no*.

(21) Every marble isn't red.
  a. None of the marbles are red.
  b. Not all of the marbles are red.

(22) Some of the marbles aren't red.
  a. Not all of the marbles are red
  b. None of the marbles are red.

(23) None of the marbles aren't red.
  a. All of the marbles are red.
  b. Some of the marbles are red.

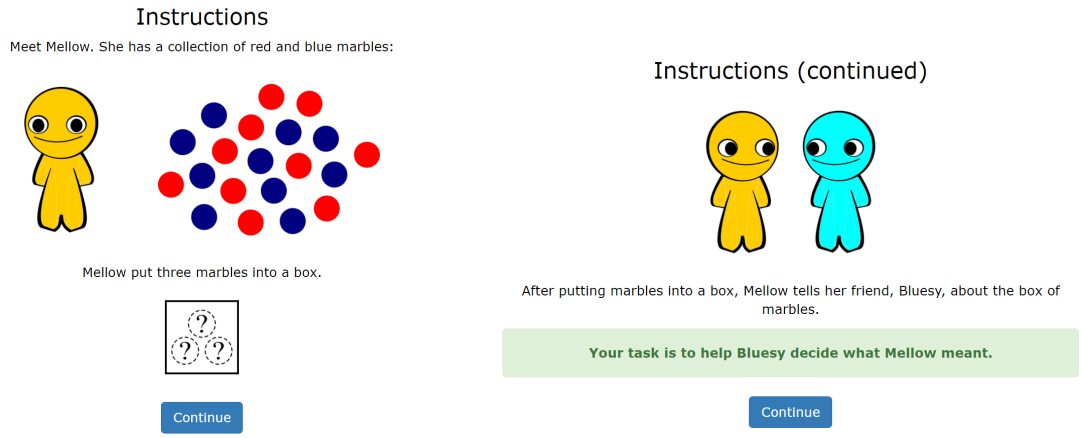
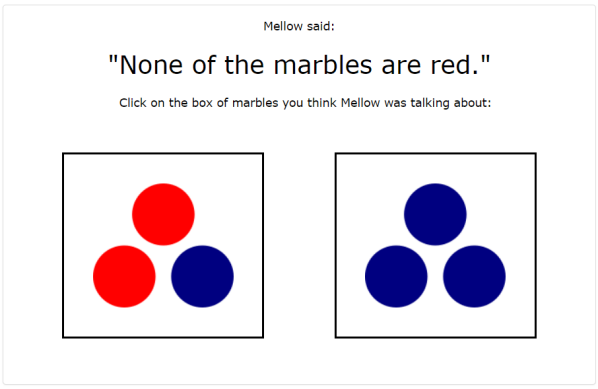
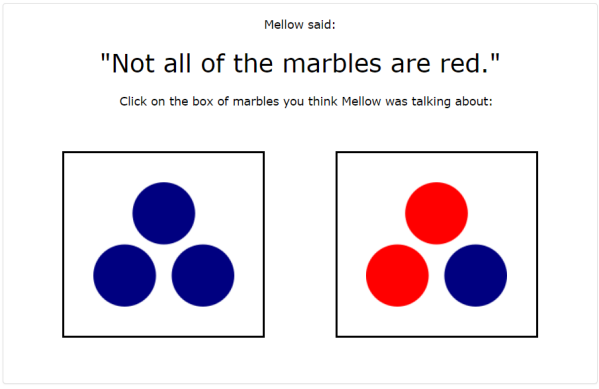


Figure 12: Instructions introducing the communication scenario used in the cross-quantifier interpretation experiments.

Results Figure 14 shows responses as the proportion of time that participants chose the inverse scope-verifying image, grouped by utterance type (ambiguous, inverse, surface) and quantifier condition. Participants chose at ceiling the image consistent with the intended scope interpretation for each of the unambiguous paraphrases: Figure 14, middle panel, shows inverse proportions near 1.0



(a) Validating surface paraphrase: as intended, participants chose at ceiling the image with three blue marbles.



(b) Validating inverse paraphrase: as intended, participants chose at ceiling the image with two red marbles.

Figure 13: Sample trials for the two scope interpretations of *every*-negation in the paraphrase validation experiment.

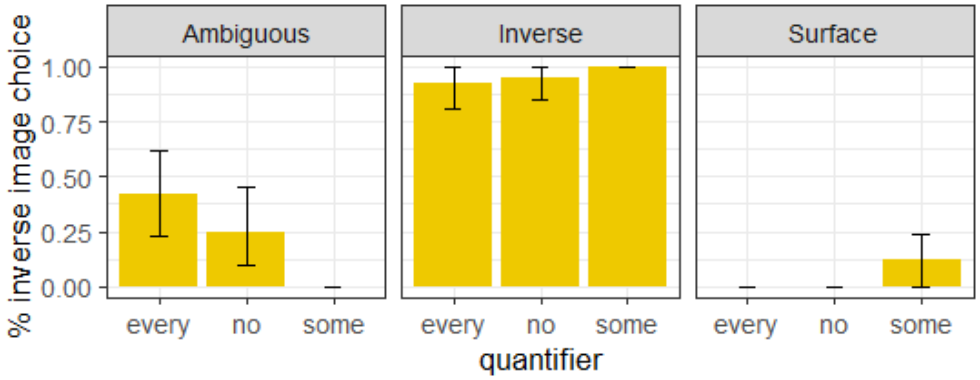


Figure 14: Paraphrase validation results. Error bars are bootstrapped 95% CIs.

for the inverse paraphrase and Figure 14, right panel, shows inverse proportions near 0.0 for the surface paraphrase. Despite the fact that *not all* and *none* can each describe a state with zero red marbles, the picture-selection data suggest that *none* and *not all* are interpreted differently (and in the way we hope) in our communication scenario.

For the potentially-ambiguous utterance, we observed a pattern (Figure 14, left panel) in line with the model predictions: *every* led to more inverse scope interpretations than *no*, which led to more inverse preference than *some*. A mixed-effects logistic regression with quantifier as a fixed effect and participant as a random intercept (maximal structure supported by the data) revealed a significant effect of quantifier ($\chi^2(2) = 17.24$, $p < .001$), but this effect was largely driven by the responses to *some*-negation, for which there were 0% inverse image choices. The difference between *every*- and *no*-negation inverse choice was not significant ($z = -1.27$, $p = .20$). We revisit this trend in the next experiment with a more sensitive measure of interpretation preferences.

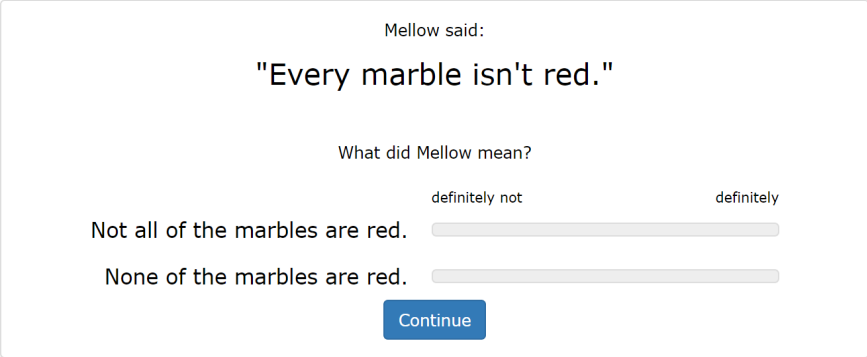


Figure 15: Sample paraphrase-endorsement trial.

5.2.2 Paraphrase endorsement

We elicited interpretations of the *every*-negation, *no*-negation, and *some*-negation utterances by asking participants to rate their validated paraphrases on a sliding scale.

Participants We recruited 60 participants with U.S. I.P. addresses through MTurk. Each received $0.50. Of the 60, we assess data from the 47 participants (32% female; mean age: 36) who indicated they understood the experiment and English was their only native language.

Design Participants saw the same communication scenario as in paraphrase validation task (Figure 12). In order to highlight the ambiguity, we presented two sliders: participants rated a slider for each of the two paraphrases of a quantifier-negation utterance (e.g., Figure 15). Note that unlike the paraphrase validation task, no images of the referents were used; further, unlike the experiment gathering annotations for the corpus, the utterances appeared on their own without linguistic context. Participants completed three trials (one for *every*, *some*, and *no*) in random order. Paraphrases were the same as those given in (21), (22), and (23).

Results Responses on the inverse and surface sliders were strongly negatively correlated across all trials ($r = -0.65$, 95% CI [-0.74, -0.55], $p < .001$), suggesting that endorsing one interpretation led to reduced endorsement for the other interpretation. A mixed-effects regression predicting surface ratings from inverse ratings (with participant as a random intercept, in order to control for participant) confirmed a robust negative relationship ($\beta = -0.72$, $p < .001$). Given that this dependency between the two sliders suggests a single underlying tradeoff between interpretations, and for the sake of consistency with the inverse-oriented focus of the rest of this paper, below, we report only the results with the inverse scope paraphrase sliders. As well, for model predictions, we follow the method used by Scontras and Goodman (2017) to only consider model predictions for one slider response.

Figure 16 shows endorsement rates (as yellow bars), grouped by quantifier, for inverse scope paraphrases, together with the fit model predictions (as dark grey bars) and unfit model predictions from Figure 9 (as pale grey bars). To assess significance, we fit a linear mixed effects model predicting the logit-transformed responses on the inverse sliders, with quantifier as a fixed effect and participant as a random intercept (as the maximal structure supported by the data); all differences were significant. Inverse preference was highest for *every*-negation, lower for *no*-negation ($\beta = -2.96$, $p < .001$), and lowest for *some*-negation ($\beta = -5.10$, $p < .001$). (An analysis based

on a combined difference score of inverse minus surface slider responses yielded the same pattern of results, with all pairwise differences between quantifiers remaining significant.) Altogether, considering the yellow bars from left to right in Figure 16: *every* allowed the most inverse interpretations (95% CI [0.65, 0.84]), *no* allowed an intermediate proportion (95% CI [0.27, 0.47]), and *some* allowed the fewest inverse scope interpretations (95% CI [0.07, 0.18]).

These behavioral results are qualitatively in line with the overall pattern of *every* vs. *no* vs. *some* interpretation preferences of the unfit model predictions, as described in Section 5.1.2 and shown by the pale grey bars in Figure 16. Inverse scope is most preferred for *every*-negation and least preferred for *some*-negation. Quantitatively, given utterance costs reflecting utterance frequencies and no other parameter fitting (maintaining minimal assumptions of $\alpha = 1$ and no expectations about the general probability of surface vs. inverse scope or the rate of marbles being red), the model is able to capture some of the pattern of average, cross-speaker interpretation preferences across quantifiers: model predictions fall just within the 95% CI for mean inverse scope probability for *every*, but overpredict the inverse scope preference for *no* and *some*.

To improve model fit, we increased the prior probability of a marble being red p_r from 0.5 to 0.67 and increased the decisiveness parameter α from 1 to 1.65, keeping utterance costs realistic and scope priors uninformative. By increasing the prior over marbles being red, we increased the degree to which the model assumed a high positive expectation. Correspondingly, the fit model is able to quantitatively match the preferred interpretations of each type of quantifier-negation utterance.[6]

5.2.3 Discussion

The results of the paraphrase endorsement task show that average interpretation preferences vary across quantifier-negation utterances that have different quantifiers: participants prefer to interpret *every*-negation with inverse scope, *some*-negation with surface scope, while *no*-negation is ambiguous but shows a slight surface scope interpretation preference. Our ambiguity resolution model, without parameter fitting beyond incorporating utterance costs reflecting utterance frequencies, successfully predicts the relative pattern of inverse scope preference across quantifier. With parameter fitting—namely, incorporating a greater high positive expectation and fitting α—we quantitatively capture the results as well.

The reason that the model, given an increased high positive expectation, successfully accounts for all three interpretation preferences remains the same as for its account of *every*-negation alone: listeners prefer the most plausible interpretation given their priors. When we increased p_r from 0.5 to 0.67, we increased the probability on the *all* world state relative to the *not all* world states, and the *none* state becomes even more unlikely. Expecting this state of affairs, listeners of *every*-negation and *some*-negation still believe it unlikely that a speaker intended the *none* interpretation and, therefore, must have meant the *not all* interpretation. The greater change is with listeners of

[6] By exploring the parameter space, we found that two changes were necessary to improve model fit: (a) increasing the salience of a high positive expectation (at decreased p_r, the model increasingly underpredicts inverse scope for *every*-negation and overpredicts inverse scope for *some*-negation); (b) given the higher prior values, α needs to increase (otherwise the model overpredicts inverse scope for *every*-negation and *no*-negation while underpredicting it for *some*-negation). The higher α value makes the speaker more strongly prefer utterances with higher communicative utility; given the higher p_r, increasing α means that these prior-driven differences in interpretation preference are more sharply reflected in the model's predictions. With these two changes to parameter values, different settings of $c(u)$, $P(w)$ and $P(i)$ do not change the qualitative results we report.

We also note that this approach of manually selecting parameters does not guarantee a globally optimal fit to the data, the way that automatic optimization methods of estimating parameters would. However, it allows us to make theoretically interpretable changes to the model, link parameter changes to qualitative predictions, and assess which changes were necessary to capture the key empirical patterns.

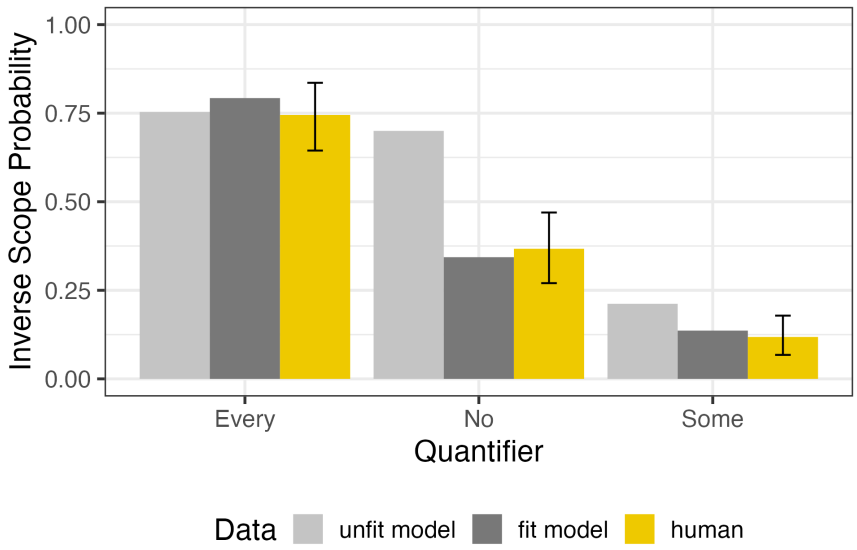


Figure 16: Results comparing model predictions and human data. Pale grey bars: Unfit model predictions for L_1 marginal distribution over interpretation i (the same as in Figure 9) with $p_r = 0.5$, utterance costs based on utterance frequencies, $P(surface) = 0.5$, and $\alpha = 1$. Dark grey bars: Model predictions fit to human data for L_1 marginal distribution over interpretation i, with $p_r = 0.67$, utterance costs based on utterance frequencies, $P(surface) = 0.5$, and $\alpha = 1.65$. Yellow bars: Degree of endorsement of the inverse scope paraphrase in the paraphrase-endorsement task. Error bars are bootstrapped 95% CIs.

no-negation: now, since they believe the *all* world state more *a priori* likely than before, they put more probability on the *all* (surface scope) interpretation than they did before.

It is especially interesting that *some*-negation is almost entirely interpreted with its surface scope interpretation. *Some* has been called a positive polarity item, an expression that for the most part does not scope under negation (Szabolcsi, 2004). These modeling results offer an explanation for why *some* might behave as a positive polarity item in the first place: interpreting *some* under negation can result in an utterance that has an unlikely meaning and is therefore inefficient.

6 General Discussion

We investigated a case study of how interlocutors might rely on world expectations in context to interpret potential scope ambiguity; more broadly, we explored how ambiguity resolution can proceed when sentences that have often been thought of as difficult or ambiguous are used as communication in context. We found that one mechanism driving interpretation preferences is plausibility: listeners tend to arrive at interpretations they believe are more likely to be true.

To formally articulate our hypothesis regarding how prior beliefs help to disambiguate *every*-negation utterances, we developed a computational cognitive model of the utterance disambiguation process. In the model, listeners assume that speakers say true things (i.e., they assume cooperativity). A listener who holds a high positive expectation that makes not-all world states more likely will be therefore more likely to attribute a not-all interpretation to an otherwise-ambiguous *every*-negation utterance. The RSA framework we used was purpose-built for the task; in particular, it articulates a principled integration of information from multiple sources (e.g., world knowledge,

semantics, context) in the process of reasoning pragmatically about utterance interpretation. Additionally, it delivers quantitative predictions that can be tested against the data we collected in our studies. Finally, we take it to be a key advantage of the framework that it requires us to clarify our assumptions regarding world knowledge, context, plausibility, and semantics in order to explicitly specify model components.

We found converging evidence for the model predictions—and thereby for the hypothesis regarding plausibility that the model implements—in a series of corpus and behavioral experiments on interpretation preferences. Specifically, we found that *every*-negation utterances (e.g., *Every vote doesn't count*) indeed receive variable interpretations in naturalistic use, and that one factor accounting for some of the variation is that listeners prefer interpretations that are more plausible given their prior beliefs about the world—particularly expectations that the relevant entities have the property corresponding to the non-negated predicate. Such beliefs make the inverse not-all interpretation more likely. Correspondingly, our model predicted greater inverse scope preference as the prior probability increases for entities having the relevant property (e.g., increased probability that votes counted, or that marbles are red, or that horses jump).

Next, in an attempt to see whether our hypothesis regarding the role of plausibility in scope disambiguation generalizes beyond *every*-negation, we extended our model to additionally include *some*-negation and *no*-negation utterances. With little parameter fitting beyond linking utterance cost to utterance type frequencies in a corpus, we accurately predicted the qualitative pattern of observed interpretations of *Every marble isn't red* vs. *Some marble isn't red* vs. *No marble isn't red* in a controlled experiment. In both the human results and our model predictions, *every*-negation receives the highest proportion of inverse scope interpretations and *some*-negation receives the lowest; for each type of quantifier-negation utterance, the model is predicting that the preferred interpretation should be the one that is most plausible given beliefs about the world.

Comparing against the findings of Scontras and Pearl (2021) for truth value judgments, these results demonstrate how the pressures driving listener behavior differ in some ways from the pressures driving speaker behavior. Speakers aim to be informative—in the RSA framework, the desire gets cashed out in the speaker's utility calculus as a pressure to effect a change between the listener's prior and posterior distribution over world states, as a way of combating the cost of speaking. In other words, speakers are happy to surprise listeners. Or, in less simplistic terms, speakers prefer to avoid saying things that are too unsurprising—why go to the trouble of saying something otherwise? On the other hand, one pressure on listeners is to bring their interpretation of a potentially-ambiguous utterance in line with their existing understanding of the state of the world. Listeners use their prior knowledge of what is likely to be true to lend weight to certain interpretations over others.

More broadly, our study helps address open questions about the naturalistic use of quantifier-negation as an instance of scope ambiguity. Scope ambiguity has been the focus of many linguistic studies, as a case study of the potential through natural language to express meaning that does not directly correspond to the overt order of a surface string of words. Yet there are many open questions about its naturalistic use, including how often scope ambiguity occurs in everyday speech, whether it is actually ambiguous in context, and if there is a preferred interpretation when both potential interpretations are attested. Through our corpus study, we found that constructions with verb negation and a subject quantified by *every* are indeed attested in transcripts of conversational speech, although they are not common. We also found similar preliminary evidence for quantifier-negation utterances with *some* and *no*. Through our behavioral study, we further confirmed that all three of these constructions are potentially ambiguous, though *some*-negation is overwhelmingly interpreted with surface scope and *every*-negation is usually interpreted with inverse scope.

Although we have focused on *every*-negation, we expect the plausibility-based interpretation

pressures we identify to apply broadly in the resolution of quantifier scope ambiguities. For example, for the scopally ambiguous cases in (24) and (25), Srinivasan and Yates (2009) suggest that plausibility would be a disambiguating factor (without a computational-level description of a disambiguation mechanism). Specifically, Srinivasan and Yates write that inverse scope is more preferred for (25) than for (24), because the surface scope interpretation in (25), that a single doctor lives in all the cities, is too implausible to be likely.

(24) A kid climbed every tree.

a. **There is a single kid who climbed all the trees.** *Surface scope* (a > every)

b. Each tree was climbed by potentially different kids. *Inverse scope* (every > a)

(25) A doctor lives in every city.

a. There is a single doctor who lives in all the cities. *Surface scope* (a > every)

b. **Each city has a different doctor living there.** *Inverse scope* (every > a)

Although it remains for future work to test the preferred interpretations of constructions like these examples, we suggest that our broader hypothesis might apply: the predicted preferred interpretation is the one that is more plausible, relative to the dispreferred interpretation, because (i) listeners are more likely to arrive at an interpretation that is likely to be true, and (ii) listeners are more likely to attribute an interpretation to a speaker (and therefore arrive at that interpretation themselves) if that interpretation would have been useful to the speaker. This future work would need to specify its assumptions (e.g., about the semantics and costs of different utterances) in order to generate predictions, to avoid automatically predicting the same interpretation preferences even for utterances with truth-conditionally equivalent quantifiers like *all*, *each*, and *every*.

While this work singles out one aspect of context we believe to influence disambiguation for *every*-negation utterances (plausibility, as defined by P(w) for the pragmatic listener, and relating broadly to rational, cooperative reasoning as encoded by RSA models), other potentially-influential aspects of the context include a range of extralinguistic and linguistic factors. For instance, the available visual context might strongly disambiguate by making certain interpretations more likely; a visual manipulation of prior expectations would be an exciting direction for future work testing the model's predictions about the role of plausibility (e.g., by showing a set of marbles that are almost all red or almost all blue prior to gathering interpretation preferences about utterances like *Every marble isn't red.*) Additionally, in the local linguistic context, prosodic stress and phrasing, lexical choice, and polarity-sensitive items in the construction might influence interpretations. Disambiguation might also depend on the immediate discourse context, information structure, and discourse goals, including the Question under Discussion (QUD) that the utterance addresses. Indeed, expectations about what the QUD is are also likely to matter, as supported by Scontras and Pearl's (2021) model for truth value judgments. However, Scontras and Pearl showed that QUDs alone are unlikely to fully explain the observed truth-value judgment behavior, and that we must also take into consideration expectations about the world state—as we have done here.

Moreover, although we focused on the pragmatic factor of plausibility, an exciting prospect for future work is to investigate another pragmatic factor, utterance alternatives, and the role that reasoning about utterance alternatives might play for preferred scope interpretations. We might expect speakers in many cases to prefer unambiguous paraphrases over the ambiguous construction. For example, *No one is the same* has fewer syllables and might well be more frequently said than the *every*-negation utterance *Every person is not the same* when intended with surface scope. A listener might then reason on the basis of a manner implicature that a speaker who went to the trouble of producing the *every*-negation utterance intended its marked inverse interpretation. (Although

it's worth pointing out that this manner implicature would predict that speakers wouldn't like to use *every*-negation with surface scope, which, similarly to our findings for the role of plausibility, doesn't align with the general expectation from the prior literature that surface scope should be preferred across the board for scopally ambiguous constructions.) Notably, this reasoning is already captured in the RSA framework. In particular, our RSA model includes considerations about what else the speaker could have said, where utterances are evaluated relative to their cost and how well they convey the speaker's intended meaning. It is less obvious which principles should determine the alternatives that should be included in the space for a range of quantifier-negation utterances. Here, we followed Scontras and Pearl (2021) in including the 'say-nothing' alternative in our model, which certainly does not exhaust the possibility space. However, our approach offers clear guidance on how to incorporate additional utterance alternatives in future explorations. An account that describes the effects of reasoning about alternatives could be the basis for an account about why speakers use scope ambiguity in the first place.

Taken together, our findings are consistent with the broader view that a sentence such as *Every vote doesn't count*, on its own, has an under-determined meaning, so that listeners fill in meaning by reasoning with information such as context and communicative intent (Grice, 1975; Sperber and Wilson, 1986). Moreover, our findings accord with the prediction, based on this broader view, that spoken language used in a linguistic and social context should often be intended and interpreted with a single interpretation; that is, language in naturalistic context should show less ambiguity than the decontextualized text that we often study. Understanding and quantifying these links between context and disambiguation stands to improve theories of language use and how use interacts with linguistic structure. We have begun quantifying these links by providing an empirical characterization of how often speakers use potentially ambiguous utterances in spontaneous speech and how ambiguous those constructions really are, together with a concrete hypothesis for how disambiguation in context could proceed.

7 Conclusion

We asked how people navigate ambiguity, specifically in quantifier-negation sentences. We confirmed that such sentences are indeed ambiguous, with interpretation preferences varying across quantifiers and contexts. We further showed that some interpretation trends can be predicted with our RSA model, which identifies shared world expectations in context—specifically, high positive expectations— as a factor shaping interpretation, in concert with knowledge of language and how language is used as communication. In particular, listeners tend to favor interpretations that are more plausible.

References

C. Anderson. *The structure and real-time comprehension of quantifier scope ambiguity.* PhD thesis, Northwestern University, Evanston, IL, 2004.

F. Beghelli and T. Stowell. Distributivity and negation: The syntax of each and every. In *Ways of Scope Taking*, pages 71–107. Springer, 1997.

G. Carden. A note on conflicting idiolects. *Linguistic Inquiry*, 1(3):281–290, 1970.

G. Carden. Multiple dialects in multiple negation. *Pap. 8th Regional Meet. Chicago Ling. Soc., ed. PM Peranteau, JN Levi, GC Phares*, pages 32–40, 1972.

G. Carden. Disambiguation, favored readings, and variable rules. *New Ways of Analyzing Variation in English*, pages 171–82, 1973.

E. S. Chung and J.-A. Shin. Native and second language processing of quantifier scope ambiguity. *Second Language Research*, page 02676583221079741, 2022.

C. Clifton Jr and A. Staub. Parallelism and competition in syntactic ambiguity resolution. *Language and Linguistics Compass*, 2(2):234–250, 2008.

M. Davies. Corpus of Contemporary American English (COCA). 2015. URL `https://doi.org/10.7910/DVN/AMUDUW`.

J. Degen. Investigating the distribution of some (but not all) implicatures using corpora and web-based methods. *Semantics and Pragmatics*, 8:11–1, 2015.

J. Degen. The rational speech act framework. *Annual Review of Linguistics*, 9, 2022.

M. C. Frank and N. D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336 (6084):998–998, 2012.

L. Frazier and J. D. Fodor. The sausage machine: A new two-stage parsing model. *Cognition*, 6 (4):291–325, 1978.

N. D. Goodman and M. C. Frank. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829, 2016.

M. Grant, S. Sloggett, and B. Dillon. Processing ambiguities in attachment and pronominal reference. *Glossa: A Journal of General Linguistics*, 5(1), 2020.

H. P. Grice. Logic and conversation. In *Speech acts*, pages 41–58. Brill, 1975.

A. Gualmini, S. Hulsey, V. Hacquard, and D. Fox. The question–answer requirement for scope assignment. *Natural Language Semantics*, 16(3):205, 2008.

J. T. Heringer. Research on quantifier-negative idiolects. In *Chicago Linguistic Society*, volume 6, page 95, 1970.

M. Hilpert. Frequency: Psychological and methodological considerations. In M. Fried and K. Nikiforidou, editors, *The Cambridge Handbook of Construction Grammar*, Cambridge Handbooks in Language and Linguistics, pages 149–170. Cambridge University Press, 2025.

S. Jusoh. A study on NLP applications and ambiguity problems. *Journal of Theoretical & Applied Information Technology*, 96(6), 2018.

K. É. Kiss and J. Pafel. Quantifier scope ambiguities. *The Wiley Blackwell Companion to Syntax, Second Edition*, pages 1–36, 2017.

H. S. Kurtzman and M. C. MacDonald. Resolution of quantifier scope ambiguities. *Cognition*, 48 (3):243–279, 1993.

G. Lakoff. On generative semantics. *Semantics: An interdisciplinary reader in philosophy, linguistics and psychology*, 232:296, 1971.

J. Lidz. The scope of children's scope: Representation, parsing and learning. *Glossa: A Journal of General Linguistics*, 3(1), 2018.

M. C. MacDonald, N. J. Pearlmutter, and M. S. Seidenberg. The lexical nature of syntactic ambiguity resolution. *Psychological Review*, 101(4):676, 1994.

C. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, 1999.

R. May. *Logical form: Its structure and derivation*, volume 12. 1985.

J. Musolino. Universal grammar and the acquisition of semantic knowledge: An experimental investigation into the acquisition of quantifier-negation interaction in English. 1999.

J. Musolino and J. Lidz. The scope of isomorphism: Turning adults into children. *Language Acquisition*, 11(4):277–291, 2003.

J. Musolino and J. Lidz. Why children aren't universally successful with quantification. *Linguistics*, 44(4):817–852, 2006.

J. Musolino, S. Crain, and R. Thornton. Navigating negative quantificational space. *Linguistics*, 38(1):1–32, 2000.

A. Neukom-Hermann. *Negation, Quantification and Scope. A Corpus Study of English and German All... Not Constructions*. PhD thesis, University of Zürich, 2016.

B. Pritchett and J. Whitman. Syntactic representation and interpretive preference. *Japanese Sentence Processing*, pages 65–76, 1995.

T. Reinhart. Coreference and bound anaphora: A restatement of the anaphora questions. *Linguistics and Philosophy*, 6(1):47–88, 1983.

E. G. Ruys and Y. Winter. Quantifier scope in formal linguistics. In *Handbook of Philosophical Logic*, pages 159–225. Springer, 2011.

G. Scontras and N. D. Goodman. Resolving uncertainty in plural predication. *Cognition*, 168: 294–311, 2017.

G. Scontras and L. S. Pearl. When pragmatics matters more for truth-value judgments: An investigation of quantifier scope ambiguity. *Glossa: A Journal of General Linguistics*, 6(1), 2021.

G. Scontras, M. Polinsky, C.-Y. E. Tsai, and K. Mai. Cross-linguistic scope ambiguity: When two systems meet. *Glossa: A Journal of General Linguistics*, 2(1):1–28, 2017.

D. Sperber and D. Wilson. *Relevance: Communication and cognition*, volume 142. CiteSeer, 1986.

P. Srinivasan and A. Yates. Quantifier scope disambiguation using extracted pragmatic knowledge: Preliminary results. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1465–1474, 2009.

A. Szabolcsi. Positive polarity–negative polarity. *Natural Language & Linguistic Theory*, 22(2): 409–452, 2004.

A. Szabolcsi. Scope and binding. *Semantics: An International Handbook of Natural Language Meaning. Mouton de Gruyter*, 2011.

S. L. Tunstall. *The interpretation of quantifiers: Semantics & processing*. PhD thesis, University of Massachusetts at Amherst, 1998.

J. Viau, J. Lidz, and J. Musolino. Priming of abstract logical representations in 4-year-olds. *Language Acquisition*, 17(1-2):26–50, 2010.

T. Wasow, A. Perfors, and D. Beaver. The puzzle of ambiguity. *Morphology and the web of grammar: Essays in memory of Steven G. Lapointe*, pages 265–282, 2005.