Using world knowledge to interpret quantifier-scope ambiguity

Noa Attali[a], Gregory Scontras[a], Lisa Pearl[a], & Stefanie Wulff[b]
nattali@uci.edu, g.scontras@uci.edu, lpearl@uci.edu, swulff@ufl.edu
[a]University of California, Irvine; [b]University of Florida

Abstract

How do we navigate potential sentence ambiguity in natural language? We combine evidence from corpus analysis, behavioral experiments, and computational modeling to motivate and formalize the utterance understanding process that yields disambiguation in context. Using the case study of quantifier-negation sentences (e.g., *Every vote doesn't count*), we find that shared world expectations in context—specifically, expectations about likely states of the world—can help explain the observed interpretation variation in naturalistic data. In particular, listeners are more likely to attribute to speakers interpretations that are more likely to be true, as listeners try to align their interpretations with what they already know about the world.

Keywords: scope ambiguity, pragmatic reasoning, 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? In particular, we explore how expectations about the world in the preceding discourse context serve as a useful source of disambiguating information.

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 Schutze, 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; although structural ambiguity is sometimes not at all associated with processing difficulty; Grant et al., 2020). On a broader view, the intended meanings of most language expressions on their own are fundamentally underdetermined, and listeners perform inferences in context in order to understand what the speaker meant to say (Grice, 1975; Sperber and Wilson, 1986).

In this paper, we quantify 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 interpret a speaker's meaning such that the interpretation is more in line with what listeners believe to be true. Our approach is to converge evidence from corpus analysis of everyday speech, controlled behavioral experiments, and computational modeling.

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 operations 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) = none of the glasses are clean
  b. (not (all the glasses are clean)) = not all of the glasses are clean

Scope ambiguity 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 and Keyser, 1985; Kurtzman and MacDonald, 1993; Musolino, 1999; Szabolcsi, 2011; Kiss and Pafel, 2017). Because scope ambiguity is a broad phenomenon, 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)

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 it.

(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 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. In other words, it is far more likely that *some or all* animation studios have moved than that *none* of the animation studios have moved (i.e., ($p(some, all) > p(none)$, 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 interpretation that the *none* interpretation is true also has a low probability. One pressure on listeners' interpretations is to align with their *a priori* understanding of the world.

We investigate this intuition by using a computational cognitive model formulated within the Bayesian Rational Speech Act (*RSA*) modeling framework (Frank and Goodman, 2012; Goodman and Frank, 2016). The model assumes boundedly rational speakers who try to minimize the cost of speaking while maximizing the probability that listeners arrive at their intended interpretation, given limits on the linguistic knowledge and information available to listeners (e.g., limits on experience with certain syntactic forms). The model specifies prior expectations about the world that are skewed a particular way. It then demonstrates 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. Model predictions say that, given skewed priors, certain scope interpretations are more likely to guide a listener to the speaker's intended interpretation. 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.

The remainder of this paper is organized as follows. First we present a quick overview of the large literature and our theoretical foundations. Then, using naturalistic examples from English corpora and crowd-sourced annotations of their preferred interpretations in context, we investigate how *every*-negation utterances are used in everyday speech. To investigate whether the world expectations that we focus on in this paper predict the observed interpretations, we measure one expression of these world expectations in the linguistic contexts of the corpus *every*-negation utterances. We test how well this expression of world expectations predicts average inverse scope preference per item and even per individual judgment. More generally to account for interpretation variation, we describe the utility of world-expectations-as-skewed-priors in a computational model that interprets *every*-negation utterances. We test analytically how the model predictions for which interpretation is preferred depends on this kind of interlocutor expectations about the world. Finally, we evaluate whether our utterance disambiguation model can generalize to correctly predicting interpretations of *different* cases of scope ambiguity in a controlled behavioral experiment. We conclude 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.

2 The role of context 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

[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

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 and Keyser, 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 (the method of collecting scope judgments is unclear), and Neukom-Hermann (2016) find 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 the first author (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?

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.*

terms for inverse scope are *nonisomorphic*, *indirect*, *NEG-Q*, or *narrow scope of the first operator.*

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. Below, we describe one of the trends we identify in the literature for the contexts that associate with *every*-negation and *all*-negation inverse scope interpretations. These kinds of contexts may facilitate inverse scope preference for children and adults; 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 experiments without context because they are often used in just these inverse-facilitating contexts, and adult speakers of English know that.

2.2 A kind of world expectation in context: High positive expectations

In the literature on scope ambiguity, one aspect of the context that consistently appears to facilitate the inverse scope preference for quantifier-negation utterances with the 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. The expectation could be held locally by interlocutors in a specific conversation, or it might correspond to more global beliefs (as a kind of world knowledge). 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 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.

One hypothesis about how high positive expectations (as a form of world expectations) influence interpretation preferences is formalized in a proposal by Scontras and Pearl (2021). They use an RSA model to articulate the cognitive process that yields observed experimental behavior for scopally-ambiguous utterances, as a way of accounting for truth value judgment patterns for quantifier-negation. In the 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 knows 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; in this study, we address this question with our own model that builds on Scontras and Pearl's. Still, the original model demonstrates the following: the more that interlocutors place a high prior probability on the *all* world state (which is a way for interlocutors to hold what we call a high positive expectation), the more that the remaining *not all* world states are *a priori* unlikely, and the more that the potentially ambiguous *every*-negation utterance becomes a highly informative way of conveying that these otherwise unlikely *not all* world states are in fact true, leading speakers to be more willing to use *every*-negation as a description of an inverse-verifying scenario. 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.

Taking stock of the existing literature, 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 a preference for inverse scope interpretations of *every*-negation and *all*-negation is a high positive expectation in the preceding context. And, one hypothesis for how high positive expectations affect behavior is articulated in Scontras and Pearl's (2021) RSA model of disambiguation.

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, but one has a small sample size (Musolino et al., 2000) and the other is based primarily on written language (Neukom-Hermann, 2016) and relied on the primary researcher to determine the intended scope interpretation. Exactly because there's 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.

With a clearer view of interpretation preferences in naturalistic speech, a further open question is whether interpretation patterns of naturalistic *every*-negation are accounted for by an RSA model of disambiguation—building on Scontras and Pearl's (2021) finding that an RSA model can account for past empirical interpretation patterns in truth value judgment studies. In exploring this question, we additionally ask what mechanisms drive the RSA model's predictions for preferred interpretations, and how well the model can account for quantifier-negation utterances more broadly.

3 Variation and high positive expectations in a corpus

To better understand interpretation preferences and the variation that people encounter, we investigated *every*-negation utterances that people produce and interpret in everyday conversation. First, we created a corpus of naturalistic *every*-negation uses by mining all *every*-negation occurrences from a corpus of conversation transcripts (Section 3.1). Then, in an experiment, we asked naive participants to indicate their scope interpretations for the *every*-negation uses occurring in their immediate contexts (Section 3.2). Finally, we explored 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 3.3).

3.1 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.

3.2 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.

3.2.1 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).

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 1. 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 1, the potentially-ambiguous clause is *Everyone does not need to establish credit by taking out a credit card*, the preceding context is *But it's helping them ...*, and the following context is *Establish credit by ...*.

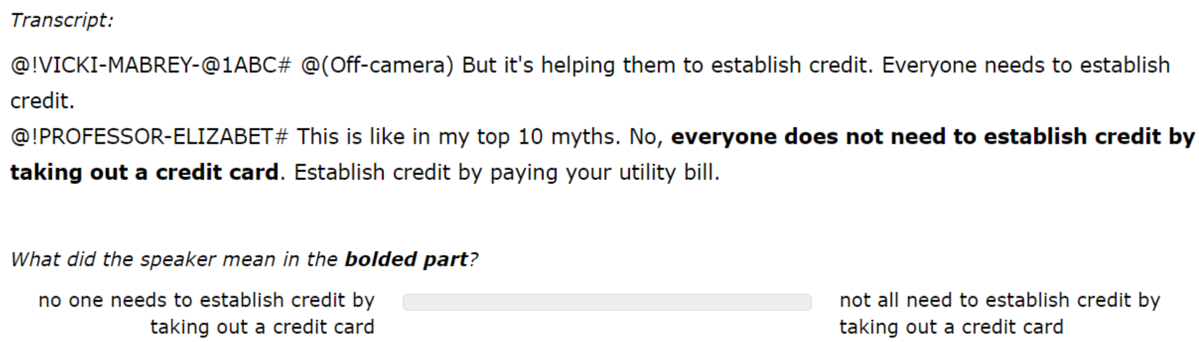


Figure 1: Sample paraphrase-endorsement trial from the corpus annotation of *every*-negation utterances.

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 1, 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 1). 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 (7), and the inverse scope-disambiguating control item is in (8). 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.

(7) 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 5.1 below.

(8) 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).

3.2.2 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 2, 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 2 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 2 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.

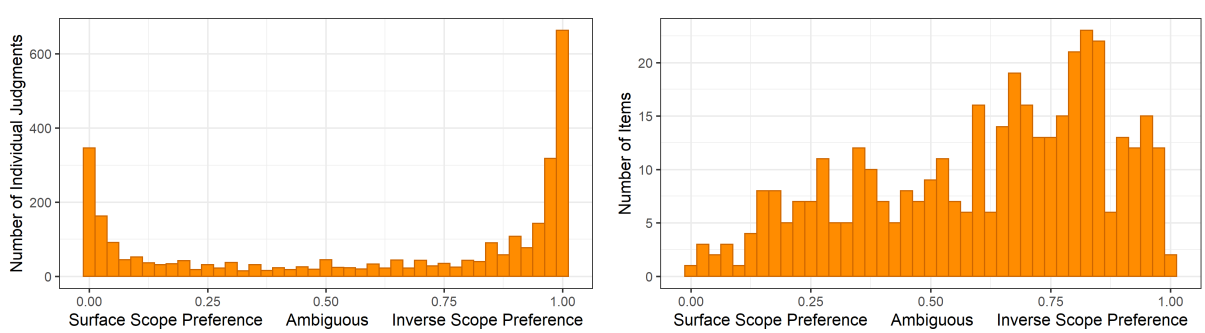


Figure 2: Individual scope interpretations (left panel) and mean interpretations per item (right panel) from the *every*-negation corpus analysis.

Figure 3 shows examples of four attested interpretation patterns: a strong surface scope preference for the item in (9) (top slider; mean response ≈ 0), a strong inverse scope preference for the item in (10) (second slider; mean response ≈ 1), and the two forms of true ambiguity (mean response ≈ 0.5). In (11) (third slider in Figure 3), we see high cross-rater disagreement, and in (12) (fourth slider in Figure 3) 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 2.

(9) (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)*

(10) 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)*

(11) 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)*

(12) 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.

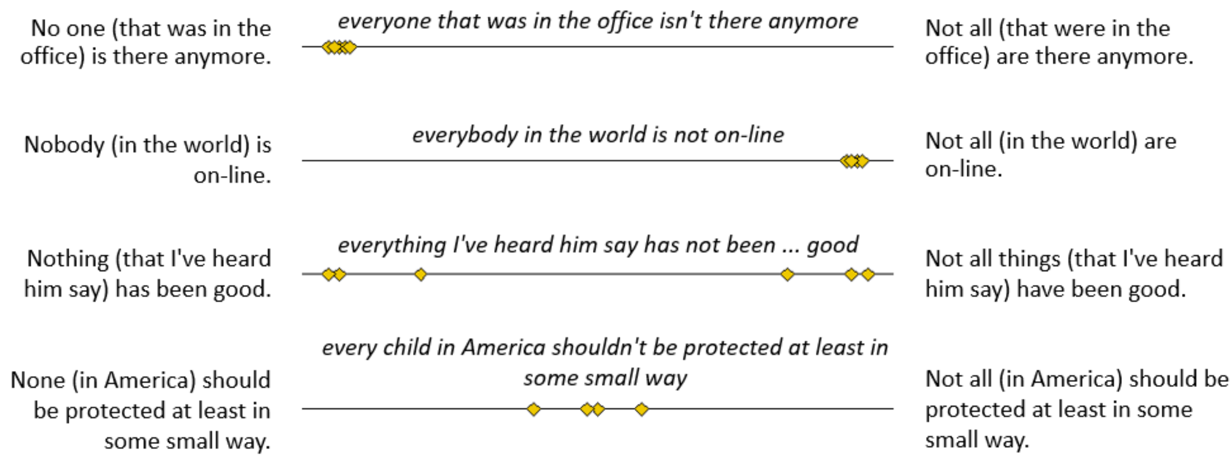


Figure 3: Individual interpretations of (9) (top slider), (10) (second slider), (11) (third slider), and (12) (fourth slider). In this figure, the horizontal line represents the length of a slider and each yellow diamond represents an individual judgment. The responses for these four items demonstrate 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).

3.2.3 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 high positive expectations can help us make sense of some of the variability in our annotated corpus.

3.3 High positive expectations in the corpus

High positive expectations may help account for the variation in interpretation preferences for *every*-negation utterances in the speech corpus. Specifically, we hypothesized that an item was more likely to receive an inverse scope interpretation in a context containing a high positive expectation. To test this hypothesis, we looked for strong 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.

3.3.1 Coding for high 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).

For an automatic, more objective, and scalable measure of the expression, we calculated the degree of lexical overlap between the preceding linguistic context and a string representing the positive expectation (*pos_exp*). For each item (e.g., *Every vote doesn't count*), we first coded *pos_exp* as the potentially-ambiguous clause without negation (e.g., *Every vote does count*). We then coded for the extent to which the *pos_exp* appeared in the preceding context as the longest common substring similarity (*LCS*; Needleman and Wunsch, 1970) between each preceding context string c and *pos_exp* pair, calculated using the R `stringdist` package (van der Loo, 2014).

Each LCS was equal to the longest sequence formed by pairing words from the preceding context string c and *pos_exp*, while keeping their order intact; the dissimilarity $d_{lcs}(c, pos\_exp)$ was then the number of unpaired words left over in both strings. Thus, dissimilarity ranges from 0 (completely similar) to the total words W in both strings combined (completely dissimilar), where $W = length(c) + length(pos\_exp)$. $d_{lcs}(c, pos\_exp)$ can be defined recursively as in (8) for different relative lengths of the two strings to be matched against:

$$d_{lcs}(c, pos\_exp) = \begin{cases} 0, & \text{if } length(c) = \varepsilon, \\ d_{lcs}(c_{1:length(c)-1}, pos\_exp_{1:length(pos\_exp)-1}), & \text{if } length(c) = length(pos\_exp), \\ 1 + min\{d_{lcs}(c_{1:length(c)-1}, pos\_exp), & \\ \qquad d_{lcs}(c, pos\_exp_{1:length(pos\_exp)-1})\}, & \text{otherwise.} \end{cases} \tag{8}$$

There are three possible outcomes for $d_{lcs}(c, pos\_exp)$, as described in equation 8. First, the value is trivially 0 for empty strings (ϵ) (8, line 1). Alternatively, the value is based on pairing each word from both strings if the two strings have equal length (8, line 2: $length(c) = length(pos\_exp)$). For example, see the first two examples in Table 1; in these two examples, $d_{lcs}(c, pos\_exp)$ is equal to the number of unpaired words left over in both strings. Third, the value is based on the minimum LCS-distance that can be obtained from pairing all the words from the shorter string to an equal number of words from the longer string (8, line 3: otherwise). For example, see the third example in Table 1; here, d_{lcs}= -2 because all four words in *pos_exp* would pair to *every vote does count* in the context, and leave unpaired the two words *I believe*.

Finally, we calculate LCS *similarity* as negative dissimilarity: $-d_{lcs}(c, pos\_exp)$. LCS similarity ranges from 0 to -W (i.e., the total number of words in the context c and *pos_exp*), with values closer to zero indicating more lexical overlap. Values closer to zero indicate a greater similarity between the context and the high positive expectation linguistic string, and so represent a higher probability that the context contained a linguistic string transparently encoding a strong expression

Preceding context c	pos_exp	$d_{lcs}(c, pos\_exp)$
Every vote does count.	Every vote does count.	0
What is going on?	Every vote does count.	-8
I believe every vote does count.	Every vote does count.	-2

Table 1: Automatically measuring the extent to which the preceding context contains an expression of a high positive expectation. The measure, $d_{lcs}(c, pos\_exp)$, is shown for different sample contexts c of the quantifier-negation utterance *Every vote doesn't count*, for which the high positive expectation pos_exp is *Every vote does count.*

of a high positive expectation.

3.3.2 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 yield inverse interpretations.

Automatic results. We used the continuous LCS-based measure $-d_{lcs}$ to assess if a high positive expectation expression predicts an inverse scope preference per item, and ran a linear mixed effects model predicting logit-transformed mean item responses by $-d_{lcs}$ (representing LCS similarity) with random intercepts for participants. To determine whether a high positive expectation captures individual judgment variation above and beyond mean item-level variation, we used a separate model to predict logit-transformed individual item responses by LCS similarity, with random intercepts for participants and items. Both models found that LCS similarity was a significant predictor of an inverse scope preference ($p < .001$ in both). That being said, the relationship is noisy, as Figure 4 shows for mean item responses, with a marginal $R^2 = 0.024$.

Interestingly, only expressions of high positive expectations that precede—but not follow—the ambiguous utterance reliably predict an inverse scope preference, as Figure 5 shows. More specifically, a version of both models that calculated LCS similarity using overlap with the following (rather than preceding) context found LCS similarity of the following context not to be a significant predictor of either item-level or judgment-level interpretations.

To see a concrete example of this relationship, again consider our original example of an item containing a high positive expectation in (13). Here, participants indeed judged on average that inverse scope was more probable (mean rating 0.70 out of 1 on the basis of seven ratings), and our automatic method identified its context to have a relatively high probability of containing a high positive expectation (LCS similarity -57).

(13) @!KING Annette, Louisiana, hello
@!CALLER Hi. My question for Mr. Eisner was, MGM is one of my favorite places in

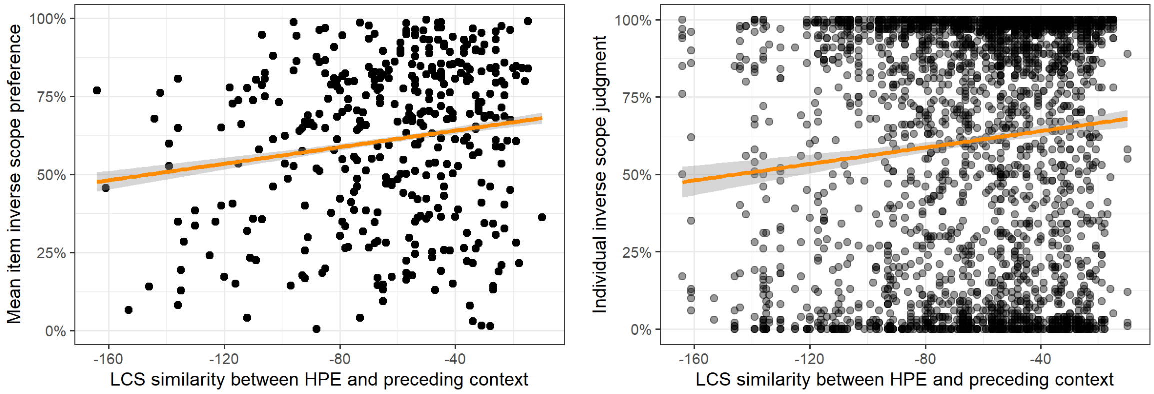


Figure 4: Preceding expression of a high positive expectation and inverse scope preference, for average item judgments and individual judgments.

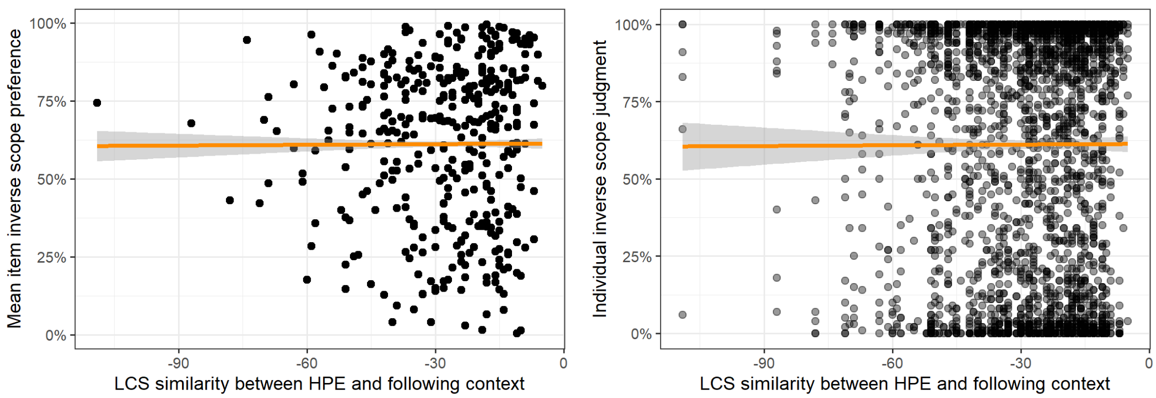


Figure 5: Following expression of a high positive expectation and inverse scope preference, for average item judgments and individual judgments.

> 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.

3.3.3 Discussion

We find that scope interpretation preferences of an individual use of *every*-negation from the corpus depend in part on whether its preceding (but not following) context expresses a strong version of a high positive expectation. In particular, high positive expectations expressed in this way correlate with stronger preferences for the inverse scope interpretation. These results align with previous modeling results (Scontras and Pearl, 2021) and pragmatically-oriented proposals from truth-value judgment studies that support the felicity of *every*-negation in context (Gualmini et al., 2008). In general, the results align with the trend we identified previously in the literature, where high positive expectations facilitate inverse scope preference for quantifier-negation with universal quantifiers.

The correlation we observe is a modest one, which may be due to our method of identifying a high positive expectation in context. In particular, the LCS similarity measure for high positive expectations likely underestimates the presence of high positive expectations. The measure looks for transparent linguistic encodings of a strong version of the belief, but of course world expectations do not have to be encoded linguistically, encoded transparently, or encoded nearby even if they are linguistically encoded. Even given our restriction to a particular form of overtly expressed world knowledge in the preceding three sentences, LCS similarity potentially underestimates the presence of a high positive expectation for several reasons. First, it is affected by context length, such that LCS similarity is lower for longer contexts even if those contexts contain a clear high positive expectation. Second, LCS similarity looks for a high positive expectation based on the exact lexical items in the *every*-negation utterance. For instance, it would identify the high positive expectation in the context *Every vote does count* for the *every*-negation utterance *Every vote doesn't count*; yet it would miss the same expectation in the context *All votes should matter* because the individual lexical items differ (*every* vs. *all*, *count* vs. *matter*). This rigidity of LCS similarity as a measure of context-sentence overlap could be a source of the noisiness evidenced in Figure 4.

Still, the advantage of LCS similarity is that it provides an automatic continuous measure to improve our analysis of larger-scale data. Here, it allowed us to consider the potential linear relationship between the extent of high positive expectation expression and the extent of an inverse scope preference.

4 Modeling scope interpretations

To better understand how speakers 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). 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 know speakers usually say things that are true and informative).

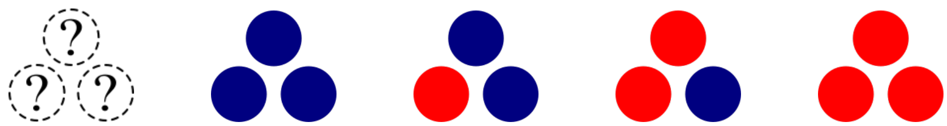
Below, we describe our model and show how it offers an explanation for the role of high positive expectations. We then extend the model in order to test the mechanism that it proposes more generally.

4.1 Modeling how high positive expectations affect *every*-negation interpretations

To account for why a high positive expectation might facilitate inverse scope preference for *every*-negation, we implement a model of utterance disambiguation for *every*-negation utterances. Our model adapts and extends an 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? We vary the extent to which the model assumes a high positive expectation and show that the predicted interpretation preference changes in a way that matches the corpus annotation results: the greater the high positive expectation, the greater the predicted inverse scope preference for *every*-negation.

4.1.1 Model articulation

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 6). In this scenario, a speaker tries to communicate the number of red marbles to a listener. The speaker can choose to say the potentially-ambiguous *every*-negation utterance (*Every marble isn't red*) or say nothing: U =\{*every*-negation, null\}.



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

Figure 6: 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 (14), 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 $Bool = \{$`true`, `false`$\}$. So, *every*-negation maps world 0 to `true` under surface scope and worlds 0, 1, and 2 (i.e., w$\neq$3) to `true` under inverse scope. The *null* utterance does not rule out any world states, mapping all of them to `true`.

(14) Utterance semantics $\llbracket u \rrbracket^i$:

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

b. $\llbracket \textit{every}\text{-negation} \rrbracket^{inverse} = \lambda \text{w. w} \neq 3$ (i.e., 'not all')

c. $\llbracket null \rrbracket = \lambda$w. `true`

Given the above-specified model universe, the RSA model describes 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 $\llbracket u \rrbracket^i$ from (14). The model implements this reasoning as a filter on the possible world states $\delta_{\llbracket u \rrbracket^i(w)}$, which returns 1 when $\llbracket u \rrbracket^i(w)$ 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 $\delta_{\llbracket u \rrbracket^i(w)}$ (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.

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

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 (α >1), smoothed away (α <1), or perceived as is (α=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.

4.1.2 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 plausibility, 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 world prior $P(w)$, which can implement a high positive expectation. To test whether high positive expectations help the model accurately predict 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).

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

4.1.3 Results

We consider the model's prediction for pragmatic listener L_1's marginal distribution over scope interpretations for the *every*-negation utterance. Figure 7 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.

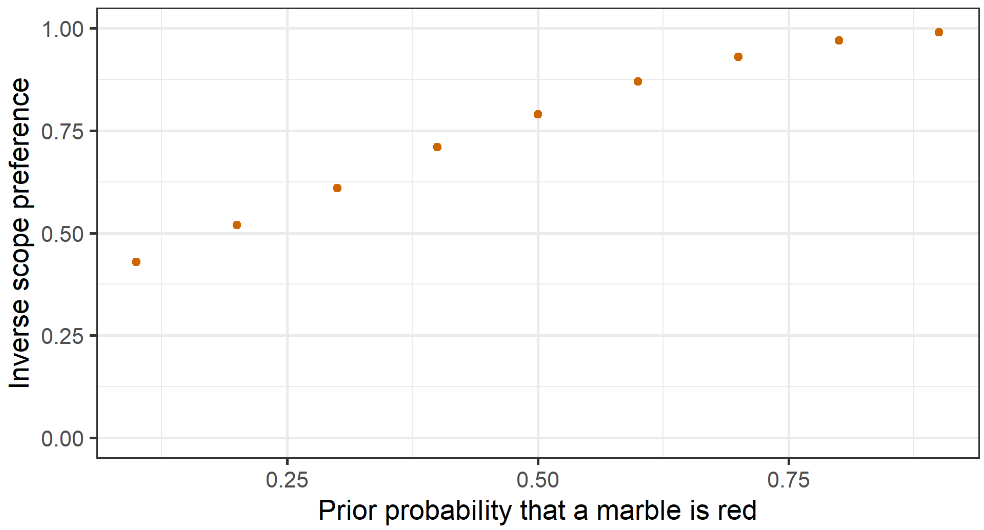


Figure 7: Predicted inverse scope preference for *every*-negation given the model's prior belief p_r that a model is a red. As the probability that each marble is red rises, the extent to which there is a high positive expectation rises, and the predicted inverse scope preference also rises.

4.1.4 Discussion

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 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*.

Note that this reasoning underlying the model predictions for L_1 listener behavior (such as we would see in a paraphrase endorsement task) 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.

4.2 Extending the model to different quantifiers

Our model of scope ambiguity resolution demonstrates how high positive expectations can explain some of the observed interpretation variation for *every*-negation utterances in context. The model's mechanism of ambiguity resolution involving world priors is meant to be general, so we turn next to assessing if 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 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.

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

All other aspects of the model articulation remain the same.

4.2.2 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. To estimate appropriate values, we used the methodology described in Section 3.1 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) $= \frac{1}{\frac{50}{390+2947+50}} = 67.741$.

4.2.3 Extended model predictions for scope interpretation preferences

Figure 8 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 4.1.4. 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 9). 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). The listener reasons that the utterance

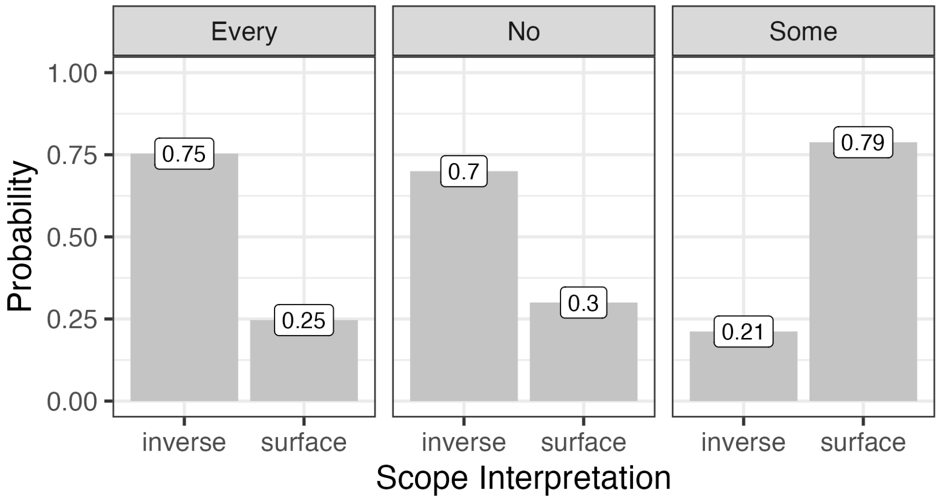


Figure 8: L_1 marginal probability distribution over scope interpretations, when we only assume relative utterance costs that reflect their relative frequencies of use in spontaneous speech (i.e., the rare *no*-negation is highly costly, *every*-negation moderately costly, and the relatively most common *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 $p_r = 0.5$.

is true, and so reasons that the speaker most likely intended the meaning that is more likely to be true: 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 likely to be true: 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.

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 of scope disambiguation and its mechanism of ambiguous utterance interpretation.

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

To test out model's predictions, we elicited native English speakers' average interpretation preferences for utterances with the quantifiers *every* vs. *some* vs. *no*. 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

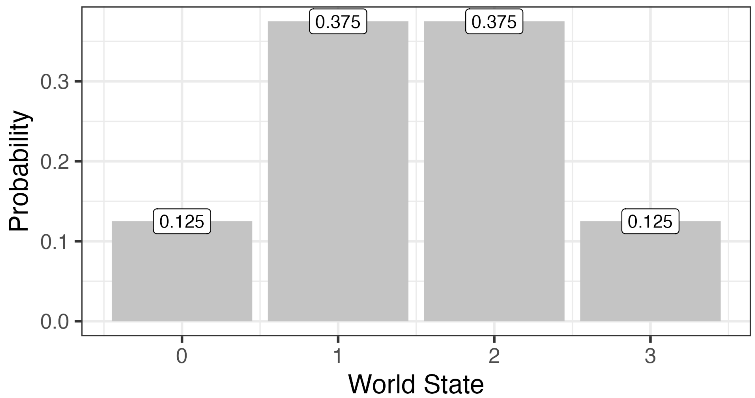


Figure 9: Prior probability distribution over world states when $p_r = 0.5$.

compatible with surface- vs. inverse-verifying scenarios.

5.1 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 10 and 11).

5.1.1 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.

5.1.2 Design

The experiment began with a scenario intended to establish that the utterances to be interpreted were communication acts (Figure 10). 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 11); 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.

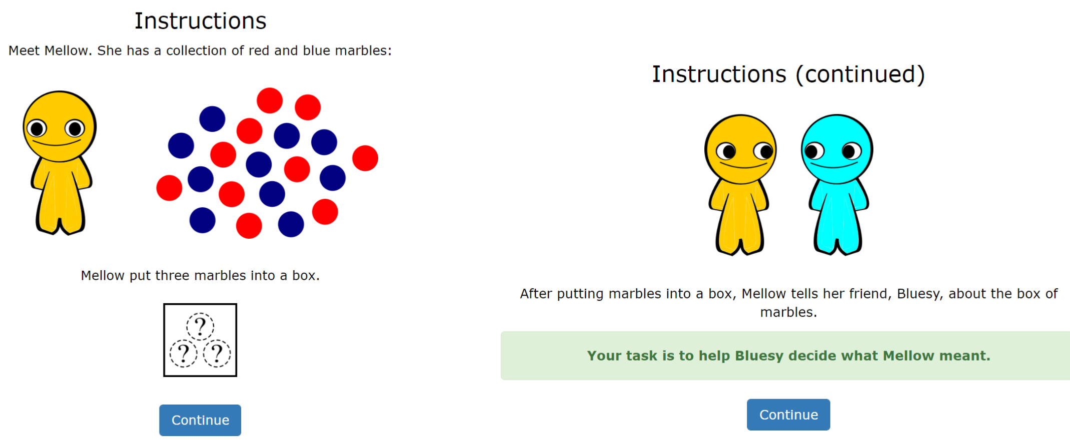
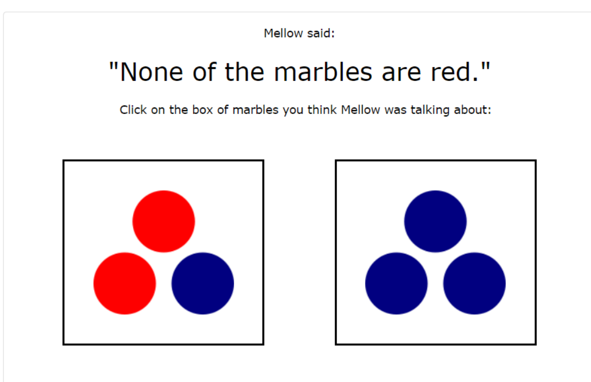
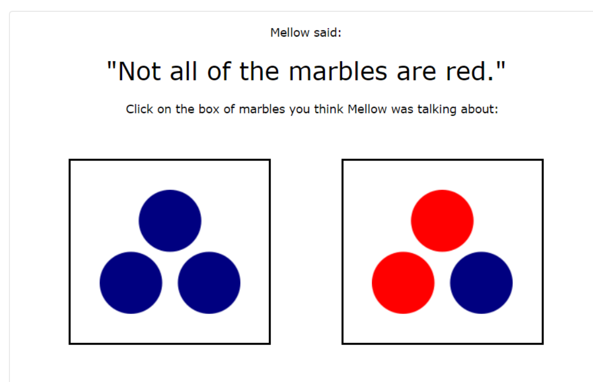


Figure 10: Instructions introducing the communication scenario used in the cross-quantifier interpretation experiments.



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

5.1.3 Results

Figure 12 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 12, middle panel, shows inverse proportions near 1.0 for the

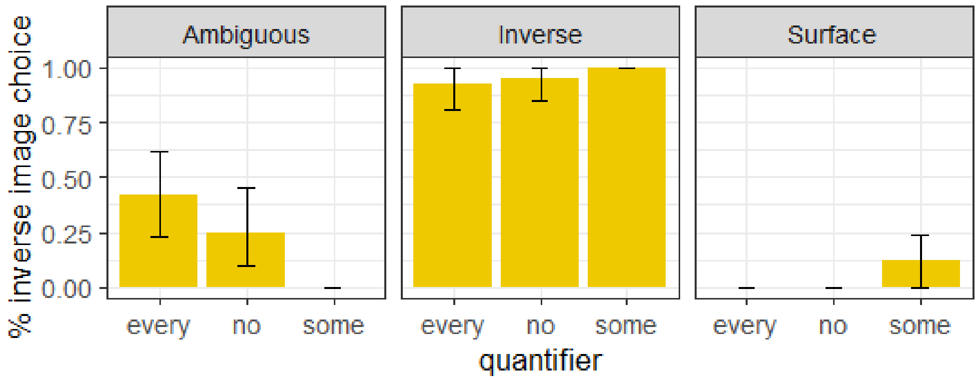


Figure 12: Paraphrase validation results. Error bars are bootstrapped 95% CIs.

inverse paraphrase and Figure 12, 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 found a non-significant trend (Figure 12, 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*. We revisit this trend in the next experiment with a more sensitive measure of interpretation preferences.

5.2 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.

5.2.1 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.

5.2.2 Design

Participants saw the same communication scenario as in Experiment 1 (Figure 10). 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 13). Note that unlike the reference task experiment, 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).

5.2.3 Results

Below, we report only the results with the inverse scope paraphrase sliders. For model predictions, we follow the method used by Scontras and Goodman (2017) to only consider model predictions for one slider response. For results in general, we found that the slider responses per item were

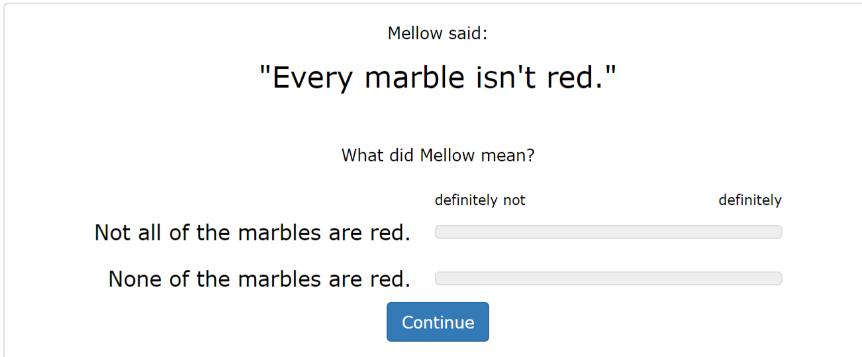


Figure 13: Sample paraphrase-endorsement trial.

negatively correlated (correlation between surface vs. inverse slider decision for *every*: -0.51; *no*: -0.40; *some*: -0.67), suggesting that endorsing one interpretation led to reduced endorsement for the other interpretation.

Figure 14 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 8 (as pale grey bars). To assess significance, we fit linear mixed effects models predicting the logit-transformed responses on each of the sliders by quantifier, with random intercepts for participant; all differences were significant. Considering the yellow bars from left to right in Figure 14: *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 4.1.3 and shown by the pale grey bars in Figure 14. Inverse scope is most preferred for *every*-negation and least preferred for *some*-negation. More specifically, 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*. The place where the unfit model qualitatively predicts the wrong preference is for *no*-negation: the unfit model predicts an inverse scope preference for *no*-negation, but the behavioral results show that *no*-negation is highly ambiguous, with a slight preference for its surface scope, *all* meaning.

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.[4]

[4]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). 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.

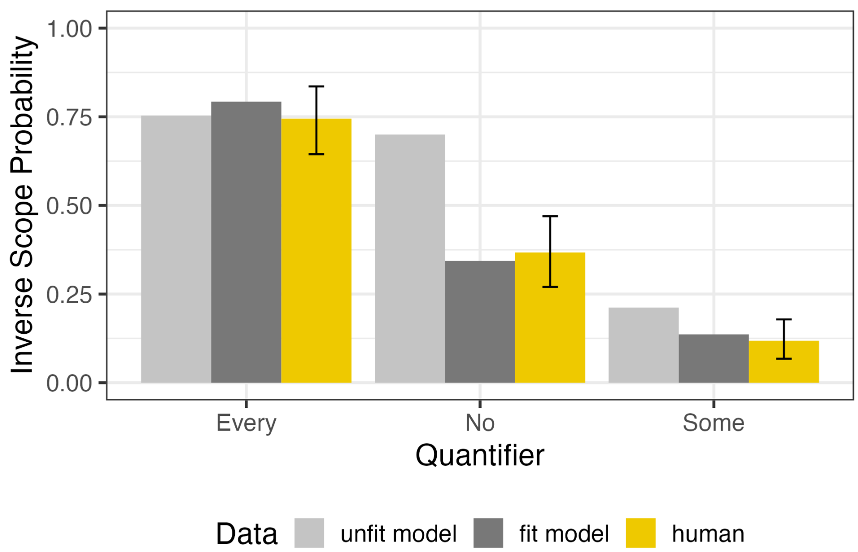


Figure 14: Results comparing model predictions and human data. Pale grey bars: Unfit model predictions for L_1 marginal distribution over interpretation i (the same as in Figure 8) with $p_r = 0.5$, utterance costs based on utterance frequencies, $P(surface) = 0.5$, and $\alpha = 1$. Dark grey bars: 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.

5.2.4 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 likely 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 *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 that listeners will try to align their interpretation with what they already believe to be true of the world.

Specifically, we found that quantifier-negation utterances (e.g., *Every vote doesn't count*) indeed receive variable interpretations, but one factor accounting for some of the variation is that listeners will prefer certain interpretations given certain skewed priors about the world. We focused on a skewed prior that we identify from the empirical literature on quantifier-negation, which we called *high positive expectations*: the belief that the relevant entities have the property corresponding to the non-negated predicate. Through an RSA model, we described how high positive expectations make more likely the inverse scope interpretation of *every*-negation, the surface scope interpretation of *some*-negation, and (slightly) the surface scope interpretation of *no*-negation. It is because listeners reason that speakers say things that are true, and high positive expectations lend relatively greater weight to world states that are compatible with these three interpretations of these utterance types. In particular, the model that is fit to human behavior provides an articulated mechanism that generates the preferred interpretations we observe for the quantifiers we investigated. A key component is that listeners expect that the *none* world state is unlikely, that the *all* world state is somewhat likely, and the *some but not all* states are most likely. With this expectation in mind, listeners then reason that speakers must have intended the *not all* rather than the *none* scope interpretations of *every*-negation and *some*-negation, and that speakers were slightly more likely to intend the *all* rather than the *some* interpretation of *no*-negation.

We found converging evidence for the model predictions in a series of behavioral experiments on interpretation preferences. First, for interpretations of natural *every*-negation utterances in context gathered from a corpus of TV and radio speech, the model successfully predicted the connection that we found in the corpus between high positive expectations and inverse scope preference. More specifically, one pragmatic factor that predicted corpus inverse scope preference was a preceding linguistic expression of a strong version of a high positive expectation (e.g., *Every vote does count* before *Every vote doesn't count*). 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).

Notably, with little parameter fitting beyond linking utterance cost to utterance type frequencies in a corpus, we predicted observed variation in several quantifier-negation combinations. We found that the model 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. *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 likely to be true. This finding demonstrates that our model of disambiguation can generalize beyond *every* and *all* in quantifier-negation utterances, highlighting the power of world expectations as part of an RSA model for capturing interpretation preferences.

These results further demonstrate how the pressures driving listener behavior differ in some ways from the pressures driving speaker behavior. Specifically, in the RSA literature and more broadly, it seems understood that one pressure from the speaker's perspective is to be informative—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. 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.

Future work can help to test how broadly the connection holds between high positive expectations and scope interpretations. One obvious extension of the current study would be to manipulate world expectations directly, either with linguistic or visual means, and then measure their effect on the interpretation that results. Future work could also replace LCS similarity, which may underestimate the prevalence of high positive expectation expression, with an automated measure that considers a vectorized semantic representation of meaning rather than lexical overlap between the context and a string representing the high positive expectation. A vectorized semantic measure would allow for the flexibility to recognize degrees of semantic similarity rather than categorical lexical equivalence. For example, such an approach would allow us to count *All votes should matter* as a context expressing a high positive expectation for *Every vote doesn't count* (recognizing that *all* is similar to *every* and *count* similar to *matter* in this context).

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 models of language use and how use interacts with linguistic structure. We have begun quantifying these links by providing an empirical characterization of how often we 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 the potential ambiguity in quantifier-negation sentences. We found that examples of this construction are indeed ambiguous, eliciting a range of average interpretation preferences depending on the quantifier and the context. Importantly, some interpretation trends can be predicted with our RSA model, which demonstrates that an important factor for interpretations is shared world expectations in context—specifically, high positive expectations—in concert with knowledge of language and how language is used as communication. Listeners are more likely to attribute interpretations to speakers that are more likely to be true.

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.

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.

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. Schutze. *Foundations of statistical natural language processing*. MIT press, 1999.

R. May and S. J. Keyser. *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.

S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970.

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

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.

M. P. van der Loo. The stringdist Package for Approximate String Matching. *The R Journal*, 6(1): 111–122, 2014. doi: 10.32614/RJ-2014-011. URL `https://doi.org/10.32614/RJ-2014-011`.

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.