2021. *Proc Ling Soc Amer* 6(1). 572–586. https://doi.org/10.3765/plsa.v6i1.4992.

Cross-linguistic scope ambiguity: An investigation of English, Spanish, and Mandarin

Yongjia Song, Abimael Hernandez Jimenez & Gregory Scontras*

Abstract. Faced with a sentence like *Every horse didn't jump over the fence* as a description of a scenario in which one out of two horses jumped, adults readily endorse the utterance as a good description, while children overwhelmingly reject it. However, systematic changes to the task setup lead to marked increases in children's endorsement rates (Musolino & Lidz 2006; Viau et al. 2010). Savinelli et al. (2017) use a computational cognitive model of utterance endorsement in truth-value judgment tasks to analytically demonstrate that both children and adults' interpretation behavior is affected by pragmatic manipulations. We test a clear prediction of these models: manipulating the conversational goal (or Question Under Discussion) should lead to clear effects on utterance endorsement. In addition to investigating the predictions for English, we also investigate Spanish and Mandarin, where the status of the relevant ambiguity may be less clear.

Keywords. scope ambiguity; questions under discussion; Rational Speech Act models; Spanish; Mandarin

1. Introduction. In English, sentences with the quantifier *every* and negation as in (1) can be interpreted in two ways, depending on the relative scope of the quantifier with respect to negation at logical form.

(1) Every horse didn't jump over the fence.
   a. Surface scope: $\forall \gg \neg$
      None of the horses jumped over the fence.
   b. Inverse scope: $\neg \gg \forall$
      Not all of the horses jumped over the fence.

Studies have found that children and adults have diverging interpretation behavior for these *every-not* sentences: where adults allow for inverse interpretations, children exhibit behavior consistent with surface interpretations (Musolino 1998; Musolino & Lidz 2006). However, the child behavior becomes markedly more adult-like as the result of changes to the context in which the sentences are interpreted. To investigate and formalize the effect of context on utterance interpretation, Savinelli et al. (2017, 2018) develop a computational cognitive model within the Rational Speech Act modeling framework (Frank & Goodman 2012; Scontras et al. electronic); the model formally articulates a hypothesis regarding how various contextual factors impact interpretation behavior.

The current paper tests a concrete prediction of Savinelli et al.'s model regarding the role that conversational goals play in utterance interpretation. We explore these interpretations in English, and also in Spanish and Mandarin. The literature on Spanish scope ambiguity is limited, but some work suggests that the status of ambiguity in Spanish may differ from English (e.g., Barberán Recalde 2017). Mandarin has been claimed to lack the ambiguity altogether (e.g., Huang 1982; Scontras et al. 2017). In our investigation, we address two questions: first,

*Authors: Yongjia Song, University of California, Irvine (yongjias@uci.edu); Abimael Hernandez Jimenez, University of Pennsylvania (abimaelh@upenn.edu); Gregory Scontras, University of California, Irvine (g.scontras@uci.edu)

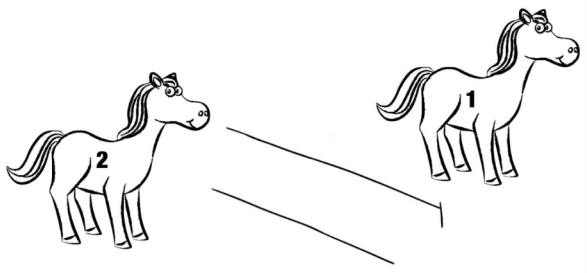

Figure 1. A not-all scenario in which one out of two horses jumps over the fence. In this scenario, the surface interpretation of the *every-not* utterance in (1) is false (it's not true that none of the horses jumped) but the inverse interpretation is true (not all of the horses jumped).

whether the relevant languages allow for scope ambiguity in *every-not* sentences like (1), and second, in cases where we believe there to be ambiguity, whether the predictions of the Savinelli et al. model hold.

After reviewing the literature on quantifier-negation scope ambiguity in English, Spanish, and Mandarin, we then describe Savinelli et al.'s model and present the predictions regarding conversational goals and utterance interpretation. To test the model predictions, we experimentally manipulate conversational goals in an utterance-endorsement task and see whether the predictions of Savinelli et al.'s model hold for English. Next, we evaluate whether a similar scope ambiguity is allowed in Spanish and Mandarin, and whether conversational goals can also affect Spanish and Mandarin speakers' interpretation behavior. Finally, we explore potential alterations to Savinelli et al.'s model that might capture our cross-linguistic results.

2. Empirical background. We begin with a brief overview of the relevant empirical observations regarding scope ambiguity in English, Spanish, and Mandarin.

2.1. ENGLISH. Most of the literature on English quantifier-negation sentences focuses on differences in interpretation behavior between children and adults. Studies commonly employ a truth-value judgment task to measure interpretation behavior (Crain & McKee 1985). In the task, participants are presented with a not-all scenario as in Figure 1. A puppet then describes the scenario, using the potentially-ambiguous *every-not* utterance in (1). Participants must then decide if the puppet's utterance offers a good way to describe the scenario. While adults readily endorse the *every-not* utterance as a good description of not-all scenarios, 5-year-old children do not, rejecting the utterance at a rate of 85-90% (Musolino 1998; Lidz & Musolino 2002; Musolino & Lidz 2006; Viau et al. 2010).

Importantly, changes to the task setup lead children to behave in a more adult-like manner, endorsing the utterance more (Musolino & Lidz 2006; Viau et al. 2010). Many of the factors affecting children's behavior in the truth-value judgment task are contextual, altering expectations about pragmatic variables. One factor hypothesized to play a central role concerns the Question Under Discussion, or QUD, which specifies the conversational goals (Roberts 2012). For an utterance to be felicitous, it must answer (at least partially) the QUD; depending on the QUD, an utterance will be more (or less) informative, which means the utterance will be more (or less) useful. In an attempt to explain differences between child and adult interpretation behavior, Gualmini and colleagues argue that children are particularly sensitive to this requirement on pragmatic felicity (Gualmini et al. 2008; Hulsey et al. 2004). In support of this claim,

Gualmini (2004) found that children's endorsement of scopally-ambiguous utterances can vary as a factor of the QUD. In their computational cognitive model of utterance endorsement in the truth-value judgment task, Savinelli et al. (2017) implement a concrete hypothesis regarding the role of QUDs in utterance endorsement behavior; we review the details of their model in Section 3 below.

2.2. SPANISH. While the literature on quantifier-negation scope ambiguity is more limited in Spanish, the existing literature suggests that Spanish may behave differently from English in its permissiveness of allowing ambiguity in the adult baseline. Barberán Recalde (2017) used a picture-matching task to assess interpretation preferences in adults and children. In the task, participants were presented with two images and asked to select the image that best paired with an utterance made by a puppet. When choosing between a not-all scenario (i.e., 2/5 success rate) and a none scenario (i.e., 0/5 success rate), both adults and children reliably selected the none scenario in response to the *every-not* utterance (adults: 83%, children: 73%). These Spanish results suggest that Spanish may be less permissive in allowing inverse interpretations for *every-not* utterances. However, it is difficult to directly compare across the different methodologies of picture selection and truth-value judgment.

2.3. MANDARIN. Unlike English, which allows quantifier scope ambiguity, Mandarin has been claimed to be much more rigid in the interpretations it allows. The Mandarin translation of (1) appears in (2). According to reports in the literature, (2) allows only an surface 'none' interpretation (Aoun & Li 1989; Huang 1982; Wu & Ionin 2019), preserving the c-command relation between the universal quantifier *mei* 'every' and negation *mei-you* 'not' at logical form.

(2) Mei-pi ma dou mei-you tiao-guo zha-lan
every-CL horse DOU NEG jump-over fence
'Every horse didn't jump over the fence.'

The experimental literature supports the theoretical claims that Mandarin does not allow inverse scope. In a truth-value judgment task, Zhou & Crain (2009) investigated the interpretation preferences of Mandarin-speaking adults. The speakers consistently rejected *every-not* sentences as in (2) as descriptions of not-all scenarios. (Curiously, Mandarin-speaking children accepted both none and not-all scenarios—a pattern opposite to that reported by Musolino & Lidz 2006 for English.) Scontras et al. (2017) investigated Mandarin and English scope relations in a different construction, namely doubly-quantified sentences as in (3) and (4).

(3) A shark attacked every pirate.
   a. Surface scope: $\exists \gg \forall$
   There was a single share that attacked multiple pirates.
   b. Inverse scope: $\forall \gg \exists$
   For each pirate, there was a (different) shark that attacked them.

(4) You yi-tiao shayu gongji-le mei-yi-ge haidao.
exist one-CL shark attack-PST every-one-CL pirate
'A/one shark attacked every pirate.'

While English speakers readily accepted both surface and inverse interpretations for sentences like (3), the Mandarin behavior was categorical: only the surface interpretations were accept-

able to participants. Importantly, the grammatical analyses meant to explain the lack of inverse interpretations in Mandarin (e.g., the Isomorphic Principle from Huang 1982) apply equally to doubly-quantified sentences as they do to quantifier-negation structures. However, it remains to be seen whether a supportive pragmatic context can lead to the possibility for inverse interpretations.

3. The model. The Rational Speech Act (RSA) modeling framework views communication between a speaker and a listener as a recursive reasoning process (Frank & Goodman 2012; Scontras et al. electronic): a speaker selects her utterance by reasoning about how a listener would interpret it; a listener interprets an utterance by reasoning about the speaker who produced it. Savinelli et al. (2017) model scope ambiguity resolution as pragmatic inference: a pragmatic listener L_1 hears a potentially-ambiguous utterance u and jointly infers the state of the world w (e.g., the number of horse that jumped over the fence) and the interpretation of the utterance i that the speaker S_1 was likely to have intended (i.e., surface vs. inverse); L_1 additionally infer the QUD q that S_1 was addressing. L_1 performs this inference by reasoning about which world state w, interpretation i, and QUD q would have been most likely to lead S_1 to produce u in the first place, weighted by the relevant prior beliefs about which world states, interpretations, and QUDs are likely in the communication scenario: $P(w)$, $P(i)$, $P(q)$. This inference is represented by the following conditional probability statement:

$$P_{L_1}(w, i, q \mid u) \propto P_{S_1}(u \mid w, i, q) \cdot P(w) \cdot P(i) \cdot P(q)$$

S_1 selects utterances u in proportion to their utility, which concerns how likely it is for u to successfully convey the intended answer x to the QUD q with a specific interpretation i to a hypothetical, naive literal listener L_0. Different QUDs will lead to different answers x depending on what the world state w is that S_1 observes.

$$P_{S_1}(u \mid w, i, q) \propto exp(\alpha \cdot log(P_{L_0}(x \mid u, i, q)))$$

The hypothetical listener L_0 interprets utterances according to their literal semantics. First, L_0 hears u with an intended i and returns a distribution of those states w compatible with the semantics of u under interpretation i:

$$P_{L_0}(w \mid u, i) \propto \delta_{[\![u]\!]^i(w)} \cdot P(w)$$

Next, L_1 uses the information about possible world states to infer the answer x to q:

$$P_{L_0}(x \mid u, i, q) \propto \sum_w \delta_{x=[\![q]\!](w)} \cdot P_{L_0}(w \mid u, i)$$

Savinelli et al. take possible world states $w \in W$ to correspond to the number of horses who jumped in a two-horse scenario, such that $W = \{0, 1, 2\}$.[1] The authors consider two possible utterances, the ambiguous *every-not* utterance and a *null* utterance that corresponds to non-endorsement in a truth-value judgment task; $U = \{\textit{every-not}, \textit{null}\}$. The utterance semantics are parameterized by i, where possible interpretations are either `surface` or `inverse`;

[1] In its original formulation, the model from Savinelli et al. (2017) investigated three-horse scenarios, such that W = {0, 1, 2, 3}. We present the two-horse case here for purposes of simplification; the change does not the affect the behavior of the model.

the interpretation of the *null* utterance is invariant to the value of i. We thus arrive at the semantics in (5).

(5) *Utterance semantics* $[\![u]\!]^i$
   a. $[\![\textit{null}]\!]^i$ = `true`
   b. $[\![\textit{every-not}]\!]^{\texttt{surface}} = \lambda \text{w. w} \neq 2$
   c. $[\![\textit{every-not}]\!]^{\texttt{inverse}} = \lambda \text{w. w} = 0$

Savinelli et al. consider three QUDs q, which serve to partition W into the possible answers to q: *how-many?* partitions W into the original four cells, corresponding to the number of horses that jumped; *all?* partitions W into two cells, one ($\{0, 1\}$) corresponding to a negative answer and the other ($\{2\}$) corresponding to a positive answer; and *none?* also partitions W into two cells corresponding to negative and positive answers (i.e., $\{1, 2\}$ for negative and $\{0\}$ for positive). To implement these partitions, the QUDs receive the semantics in (6).

(6) *QUD semantics* $[\![q]\!]$
   a. $[\![\textit{how-many?}]\!] = \lambda \text{w. w}$
   b. $[\![\textit{all?}]\!] = \lambda \text{w. w} = 2$
   c. $[\![\textit{none?}]\!] = \lambda \text{w. w} = 0$

To generate predictions about truth-value judgments, Savinelli et al. need one last ingredient: an additional inference layer corresponding to a pragmatic speaker S_2 who observes the world state w (e.g., by viewing the scenario in Figure 1) and decides whether to endorse the *every-not* utterance or to choose *null* instead. S_2 performs this inference by simulating the effect of u on L_1's marginal distribution over w:

$$P_{S_2}(u \mid w) \propto exp(log(\sum_{i,q} P_{L_1}(w, i, q \mid u))$$

By manipulating the priors over W, I, and Q, the authors are able to evaluate the effect of these factors on predicted utterance endorsement.

To generate model predictions, various free parameter values must be set. Savinelli et al. set the utility-scaling parameter α to 2.5, although similar results are obtained by setting α to 1 (i.e., no scaling; for more on the role of α in RSA, see Zaslavsky et al. 2021). The default value of the world state prior, $P(w)$, divided probability equally among the possible world states: $P(w = 0) = P(w = 1) = P(w = 2) = \frac{1}{3}$; this flat prior over world states models the idea that participants are maximally uncertain about which states are more or less likely a priori. For the prior over scope interpretations, $P(i)$, Savinelli et al. model the fact that surface interpretations are more accessible by setting $P(\texttt{surface}) = 0.7$ and $P(\texttt{inverse}) = 0.3$; the qualitative predictions we report hold also for a flat scope interpretation prior.

Savinelli et al. explore the role of QUDs in endorsement behavior by systematically manipulating the QUD prior to favor certain QUDs, while keeping the other priors at their default values.[2] The favored QUD received a prior probability of 0.9, while the two other disfavored QUDs split the remaining probability, 0.1, equally between them. This manipulation was meant to model the experimental manipulation of QUDs, privileging certain QUDs over others.

Figure 2 plots the results of this manipulation: endorsement of the ambiguous *every-not*

[2] The authors also explore the interaction among factors; see Savinelli et al. (2017) for more details.

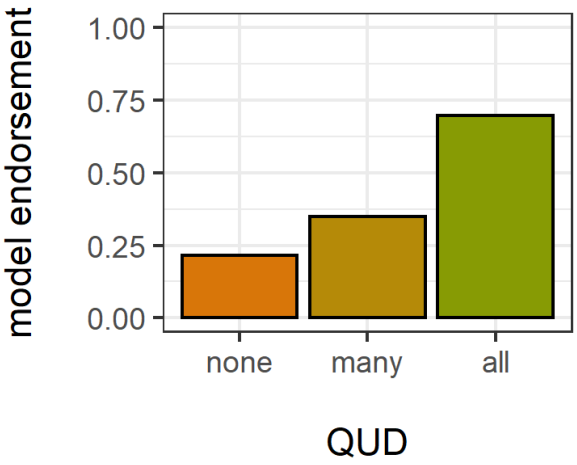

Figure 2. Model predictions of of the QUD manipulation from the Savinelli et al. (2017) model. Predictions represent the probability of S_2 choosing the ambiguous *every-not* utterance as a description of the not-all world state (i.e., $w = 1$).

utterance in a not-all world state (i.e., $w = 1$). The model predicts the most endorsement with the *all?* QUD: "Did all of the horses jump over the fence?" With this QUD, the *every-not* utterance provides a full answer (that is, 'no') under either scope interpretation; by fully resolving the *all?* QUD, the *every-not* utterance is informative and therefore useful for the speaker, which means the speaker is more likely to select it. With *none?*, the *every-not* utterance is a particularly ineffective means of conveying that it is not the case that none of the agents completed the action, so the modeled speaker in unlikely to endorse this uninformative, not useful utterance. With *how-many?* as the QUD, utterance endorsement is predicted to be intermediate between *all?* and *none?*.

4. Testing the predictions. To test the predictions of Savinelli et al.'s model of scope ambiguity resolution, we investigated the role of QUDs in utterance interpretation. We begin by directly testing the predictions in English, and then looking at Spanish to see whether the language behaves similarly. We then look at interpretation preferences in Mandarin, where we expect the relevant sentences to be unambiguous. However, it remains to be seen how rigid the scope behavior is in Mandarin, and whether that behavior may be influenced by pragmatic considerations like the QUD.

4.1. EXPERIMENT 1: ENGLISH. To test Savinelli et al.'s model predictions in English, we ran a truth-value judgment task, manipulating the QUD and measuring its effect on endorsement behavior.

Participants. We recruited 291 participants via Amazon.com's Mechanical Turk crowdsourcing service. On the basis of their response to a post-test demographics questionnaire, we identified 263 participants as native speakers of English; their data were included in the analyses reported below. All participants were compensated for completing the experiment.

Design. Participants were introduced to Shark, a character who likes to organize her storybooks. Shark's friend, Elephant, helps by reading some stories and describing them to Shark. Participants were tasked with making sure that Elephant says the right thing when describing

Progress:

Shark is trying to oraganize her storybooks about **butterflies**. She has been sorting books according to **whether none of the butterflies went to the city**.

Shark's friend, Elephant, is helping her by reading some of the books.
Here is a book that Elephant read:

This story features two butterflies, a forest, and a city. The two butterflies were deciding where to go. First, they thought about the forest, and decided to go. One butterfly did not like the forest, but the other one did. The butterfly who didn't like the forest decided to leave the forest and go to the city. The other butterfly decided to leave the forest and go home.

In order to sort the book, Shark asked:

"Did none of the butterflies go to the city?"

Elephant answered:

"Every butterfly didn't go to the city."

Is Elephant right?

definitely not definitely

Continue

Figure 3. Example *none?* trial from the English experiment.

the stories.

After this introduction, participants encountered Shark's organization scheme (i.e., her *goal structure*). In addition to stating the goal explicitly (e.g., "whether none of the butterflies went to the city" in Figure 3), participants also saw a visualization of Shark's goal structure (i.e., the labeled bins in Figure 3); goals corresponded to the relevant QUD (*none?* in Figure 3), and the goal structure visualized the possible answers to the QUD.

After seeing the goal structure, participants read the story that was read by Elephant; stories described a scenario like that depicted in Figure 1 where some but not all of the agents successfully completed an action. Next, participants observed a short dialogue between Shark and Elephant, in which Shark asks the QUD corresponding to the relevant goal structure, and Elephant responds with the *every-not* utterance. Participants then adjusted a slider to indicate whether Elephant's answer was "right" (i.e., whether they endorsed the utterance as a good description of the story), with slider endpoints labeled *definitely not* (coded as 0) and *definitely* (coded as 1).

Participants completed only a single trial where they were randomly assigned one of three QUDs: did none of the agents complete the action, did all of the agents complete the action, or how many agents completed the action; these three QUDs were the ones for which the model of Savinelli et al. (2017) makes concrete predictions. The story was chosen at random from a set of four items that differed according to their characters and actions (frogs jumping over rocks, butterflies going to the city, lions buying a cookie, dinosaurs eating bugs).

Results. Figure 4 plots endorsement rates grouped by QUD. To evaluate the effect of QUD, we fit a linear mixed effects model predicting endorsement rates by QUD, with *many?* dummy-coded as the reference level, and with random intercepts by item. Ratings for *all?* were significantly different from the ratings for *many?* ($\beta = 0.22$, $t = 3.85$, $p < 0.001$), as were the ratings for *none?* ($\beta = -0.19$, $t = -3.20$, $p < 0.01$). Thus, these data match the predictions of Savinelli et al.'s model: *all?* recieves the highest endorsement rates, followed by *many?*, and then by *none?*.

4.2. EXPERIMENT 2: SPANISH. Having found support for Savinelli et al.'s model predictions in English, we turned next to Spanish, replicating the English experiment with Spanish translations of the materials.

Participants. We recruited 400 participants through Prolific.co's crowdsourcing service.[3] We identified native Spanish speakers as those participants who indicated Spanish as their native language and who reported living in a Spanish-speaking country both before and after the age of 8 for a total of more than five years. 310 participants were identified as native speakers by these criteria; their data are included in the analyses below. All participants were compensated for their participation.

Design. This experiment was a direct translation of the English experiment, with all instructions and materials translated into Spanish. A translation of the critical *every-not* test sentence from Figure 3 appears in (7).

[3] In previous work, we had found Prolific to be a more effective tool than Mechanical Turk for selectively recruit-ing non-English speakers.

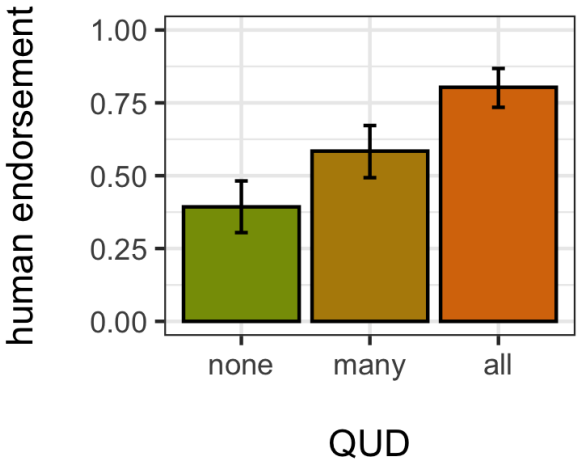

Figure 4. English speakers' endorsement of the potentially-ambiguous utterance as a good description of the not-all scenario across three QUDs (*all?, many?,* and *none?*). Error bars represent bootstrapped confidence intervals drawn from 10,000 samples of the data.

(7) Todas las mariposas no fueron a la cuidad.
all the butterflies NEG went to the city
'Every butterfly didn't go to the city' (lit. 'all the butterflies didn't go to the city')

Results. Figure 5 plots endorsement rates grouped by QUD. As with the English analysis, we fit a linear mixed effects model predicting endorsement by QUD with random intercepts by item. Compared with *many?*, responses to the *all?* QUD were significantly greater ($\beta = 0.24$, $t = 4.31$, $p < 0.001$); there was no significant difference between *many?* and *none?* ($\beta = -0.04$, $t = -0.77$, $p = 0.44$). Thus, despite a numerical trend in the predicted direction (i.e., *many?* responses received higher numerical endorsement than *none?*), Savinelli et al.'s model predictions are only partially confirmed in Spanish. In Section 5 below, we discuss possible explanations for this result.

4.3. EXPERIMENT 3: MANDARIN. Finally, we turn our sights to Mandarin. Given claims in the literature, we expect Mandarin speakers to reject *every-not* utterances as a description of not-all scenarios—a departure from English and Spanish. However, we were interested in seeing whether this behavior is modulated by the QUD in Mandarin.

Participants. We recruited 140 participants from Prolific. We identified native speakers of Mandarin as those participants who indicated Mandarin (Chinese) as their native language and who reported living in a Chinese-speaking country both before and after the age of 8 for a total of more than five years. 79 of the participants met these criteria; their responses are reported below. All participants were compensated for their participation.

Design. The experiment was identical to the English and Spanish experiments, except all materials and instructions were translated into Mandarin. A translation of the *every-not* test sentence from Figure 3 appears in (8).

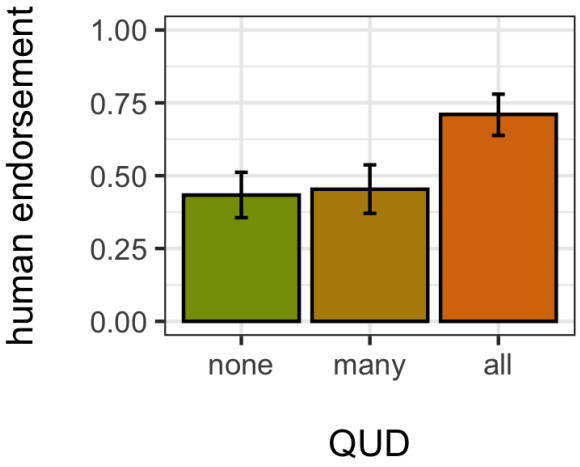

Figure 5. Spanish speakers' endorsement of the potentially-ambiguous utterance as a good description of the not-all scenario across three QUDs. Error bars represent bootstrapped confidence intervals drawn from 10,000 samples of the data.

(8) Mei-zhi hu-die dou mei-you qu cheng-shi
every-CL butterfly DOU NEG go city
'Every butterfly didn't go to the city.'

Results. Figure 6 plots Mandarin endorsement rates grouped by QUD. As with English and Spanish, we fit a linear mixed effects model predicting endorsement by QUD with random intercepts by item. Compared to the *many?* QUD, neither *all?* ($\beta = -0.12$, $t = -1.56$, $p = 0.12$) nor *none?* ($\beta = 0.01$, $t = 0.09$, $p = 0.93$) received significantly greater endorsement.

Visual inspection of the results suggests that Mandarin speakers provided lower endorsement overall compared to English and Spanish. To assess this difference, we pooled the data from all three experiments and fit a linear mixed effects model predicting endorsement by language with Mandarin dummy-coded as the reference level; the model included random by-item intercepts. As expected, English speakers provide higher endorsement than Mandarin speakers ($\beta = 0.36$, $t = 6.92$, $p < 0.001$), as do Spanish speakers ($\beta = 0.28$, $t = 3.65$, $p < 0.01$).

4.4. SUMMARY. Taken together, the results from our three experiments partially confirm the model predictions from Savinelli et al. (2017): we see a clear effect of QUD in English such that *all?* has higher rates of endorsement than *many?*, which has higher endorsement than *none?*—precisely the pattern of results predicted by Savinelli et al. This effect is partially replicated in Spanish, where *all?* has higher endorsement than the other QUDs. In Mardarin, we fail to find an effect of QUD. Comparing across languages, we see that Mandarin has overall lower rates of endorsement than the other two languages, suggesting that English and Spanish allow for ambiguity in the *every-not* utterance, while Mandarin does not. This interpretation is further supported by the absence of a QUD effect in Mandarin: contextual support (in the form of a supportive QUD) does not lead the Mandarin sentence to felicitously describe a not-all scenario.

5. Exploring model behavior. The Savinelli et al. model straightforwardly captures the behavior of our English participants. In what follows, we explore changes to the model that may

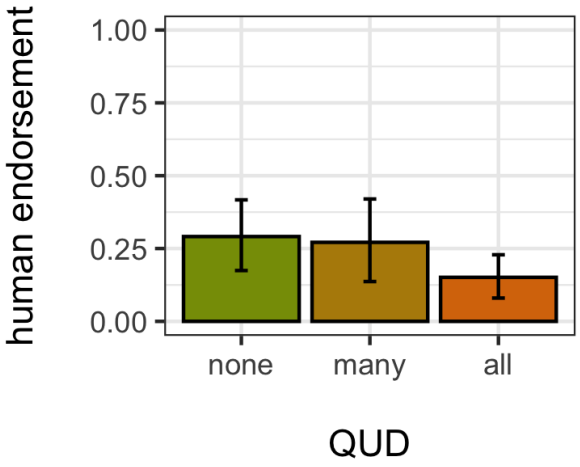

Figure 6. Mandarin speakers' endorsement of the potentially-ambiguous utterance as a good description of the not-all scenario across three QUDs. Error bars represent bootstrapped confidence intervals drawn from 10,000 samples of the data.

capture the Spanish and Mandarin behavior as well.

5.1. MODELING SPANISH. Spanish participants endorse the *every-not* utterance more with the *all?* QUD than with *many?*, as predicted by the model. However, the model also predicts a difference in endorsement rates between the *many?* and *none?* QUDs that, although present numerically in our behavioral data, does not reach significance.

One possibility is that, for some yet-to-be-determined reason, our *none?* QUD manipulation in Spanish was less effective at privileging the *none?* QUD than was the corresponding manipulation in the English experiment. In other words, perhaps Spanish participants were less swayed by the *none?* QUD in our experiment. We can implement this hypothesis in our model by changing our assumptions about the values of the QUD prior. Figure 7 plots predictions from a version of the model with the following priors:

(9) QUD priors in the modified Spanish model:
   a. *none?* favored: $P(\textit{none?}) = \mathbf{0.6}$, $P(\textit{how-many?}) = 0.2$, $P(\textit{all?}) = 0.2$
   b. *many?* favored: $P(\textit{none?}) = 0.125$, $P(\textit{how-many?}) = \mathbf{0.75}$, $P(\textit{all?}) = 0.125$
   c. *all?* favored: $P(\textit{none?}) = 0.125$, $P(\textit{how-many?}) = 0.125$, $P(\textit{all?}) = \mathbf{0.75}$

Comparing the model prediction in Figure 7 with the human behavior in Figure 5, we see that these modeling choices come close to capturing the behavioral data.

The other possibility is that the original predictions of the model in fact hold in Spanish, but our experiment was not sensitive enough to pick up on the difference between the *many?* and *none?* conditions. Recall the model predictions in Figure 2: the predicted difference between *many?* and *all?* is larger than the predicted difference between *many?* and *none?*. Numerically, this pattern is precisely what we find in the Spanish results; however, the difference between *many?* and *none?* does not reach significance. It remains to be seen whether an experiment with more power would pick up on this difference, or whether the QUD manipulation is to blame, as outlined above.

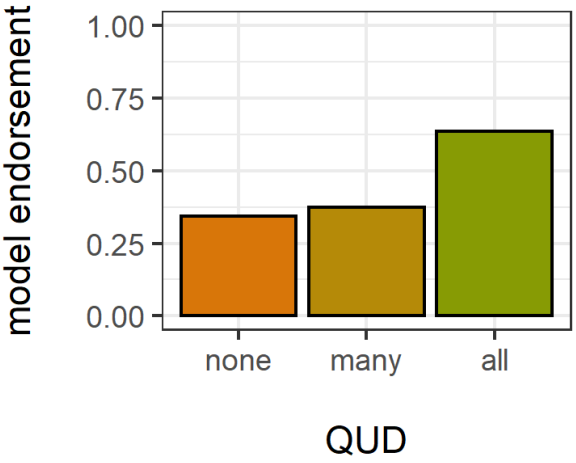

Figure 7. Predicted endorsement of the unambiguous *every-not* utterance in a not-all scenario in a model using the QUD manipulation in (9).

5.2. MODELING MANDARIN. In Mandarin, unlike in English and Spanish, we found low endorsement for the *every-not* utterance regardless of the QUD, suggesting a lack of ambiguity for *every-not* utterances in Mandarin consistent with claims in the literature. As a fist attempt at modeling this behavior, we might try altering the utterance semantics in the model so that *every-not* unambiguously receives a surface interpretation. Unfortunately, this change will not suffice to capture the Mandarin behavior. The information-theoretic pressures driving the QUD effect in the model apply also in the absence of ambiguity: endorsement rates are high for *all?* because either interpretation, surface or inverse, provides an informative answer to the QUD. Thus, even if the *every-not* utterance unambiguously receives a surface interpretation, it should still be endorsed more with *all?* (for discussion, see Chen & van Tiel 2021).

To model the Mandarin results, we need more than just an unambiguous utterance semantics. One possibility for why Mandarin speakers resist endorsing *every-not* utterances as descriptions of not-all scenarios is that there exists a salient, cheap alternative utterance for communicating the not-all meaning. In fact, Zhou & Crain (2009) mention that the *not-every* sentence in (10) serves just this role, as an unambiguous way to describe not-all scenarios in Mandarin. Importantly, English also allows this configuration of *every* and negation as an utterance alternative, but Zhou & Crain claim that the English *not-every* utterance is less common than the *every-not* utterance in (1), suggesting that *not-every* is not as cheap of an utterance alternative in English as it may be in Mandarin.

(10) bu-shi mei-pi ma dou tiao-guo zha-lan
NEG every-CL horse DOU jump-over fence
'Not every horse jumped over the fence.'

In updating the model, we can attempt to capture these facts about utterance alternatives in Mandarin by fixing the semantics of *every-not* to surface interpretation only and including the additional unambiguous *not-every* utterance alternative: $U = \{\textit{every-not}, \textit{not-every}, \textit{null}\}$; we keep the other parameters the same. Predictions of this version of the model are shown in Figure 8. While this change effectively eliminates endorsement of the *every-not* utterance

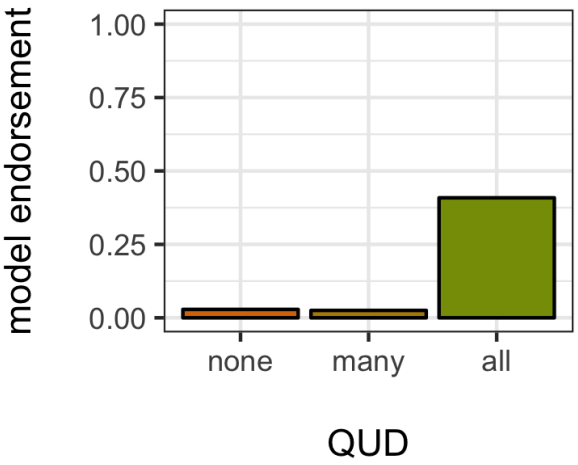

Figure 8. Predicted endorsement of the unambiguous *every-not* utterance in a not-all scenario in a model that also include an unambiguous *not-every* utterance; the QUD manipulation is implemented by assigning 90% of the prior probability mass to the privileged QUD and distributing the remaining probability across the other two QUDs.

with the *none?* and *many?* QUDs, endorsement is predicted to be substantially higher for the *all?* QUD, contrary to what we observed behaviorally. It seems, then, that the model requires some other amendment to capture the rigid scope behavior observed in Mandarin; we leave it to future work to explore model architecture options.

6. Discussion. We have found partial support for the information-theoretic modeling predictions from Savinelli et al. (2017): English participants are more likely to endorse an ambiguous *every-not* utterance as a description of a not-all scenario when that utterance serves as a better answer to the operative QUD. In the case of *all?*, the *every-not* utterance provides a full answer (i.e., 'no') under either interpretation; with *none?*, the *every-not* utterance is a particularly ineffective means of conveying that it is not the case that none of the agents completed the action. In Spanish, we find the predicted increased endorsement for *all?*, although the status of the *none?* QUD is less clear. Still, the results suggest that Spanish does allow for scope ambiguity in *every-not* sentences, and with some modifications to parameter values, Savinelli et al.'s model can predict the Spanish pattern of results. In Mandarin, the picture looks different: we find low endorsement across the board for *every-not* utterances as descriptions of not-all scenarios, and alternations to the parameter values, utterances alternatives, and semantics fail to capture the Mandarin pattern of behavior. It seems, then, that an additional mechanism is required to derive the rigid-scope interpretation behavior in Mandarin.

References

Aoun, Joseph & Yen-hui Audrey Li. 1989. Scope and constituency. *Linguistic Inquiry* 20. 141–172. https://www.jstor.org/stable/4178623.

Barberán Recalde, Tania. 2017. The interpretation of negated quantifiers in Spanish. In Chelo Vargas-Sierra (ed.), *Professional and academic discourse: An interdisciplinary perspective* (EPiC Series in Language and Linguistics, vol. 2), 181–190. Manchester, UK: EasyChair. https://doi.org/10.29007/vt9f.

Chen, Sherry Yong & Bob van Tiel. 2021. Every ambiguity isn't syntactic in nature: Testing the rational speech act model of scope ambiguity. *Proceedings of the Society for Computation in Linguistics* 4(1). 24. https://scholarworks.umass.edu/scil/vol4/iss1/24.
Crain, Stephen & Cecile McKee. 1985. The acquisition of structural restrictions on anaphora. *Proceedings of the North East Linguistic Society (NELS)* 15. 94–110.
Frank, Michael C. & Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science* 336(6084). 998–998. https://doi.org/10.1126/science.1218633.
Gualmini, Andrea. 2004. Some knowledge children don't lack. *Linguistics* 957–982. https://doi.org/10.1515/ling.2004.034.
Gualmini, Andrea, Sarah Hulsey, Valentine Hacquard & Danny Fox. 2008. The question–answer requirement for scope assignment. *Natural Language Semantics* 16(3). 205–237. https://doi.org/10.1007/s11050-008-9029-z.
Huang, Cheng-Teh James. 1982. *Logical relations in Chinese and the theory of grammar*. Cambridge, MA: MIT dissertation.
Hulsey, Sarah, Valentine Hacquard, Danny Fox & Andrea Gualmini. 2004. The question-answer requirement and scope assignment. *MIT Working Papers in Linguistics* 48. 71–90.
Lidz, Jeffrey & Julien Musolino. 2002. Children's command of quantification. *Cognition* 84(2). 113–154. https://doi.org//10.1016/S0010-0277(02)00013-6.
Musolino, Julien. 1998. *Universal Grammar and the acquisition of semantic knowledge: An experimental investigation into the acquisition of quantifier-negation interaction in English* College Park, MD: University of Maryland dissertation.
Musolino, Julien & Jeffrey Lidz. 2006. Why children aren't universally successful with quantification. *Linguistics* 44. 817–852. https://doi.org/10.1515/LING.2006.026.
Roberts, Craige. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics* 5(6). 1–69. https://doi.org/10.3765/sp.5.6.
Savinelli, K. J., Gregory Scontras & Lisa Pearl. 2017. Modeling scope ambiguity resolution as pragmatic inference: Formalizing differences in child and adult behavior. *Proceedings of the Annual Meeting of the Cognitive Science Society* 39. 3064–3069.
Savinelli, K. J., Gregory Scontras & Lisa Pearl. 2018. Exactly two things to learn from modeling scope ambiguity resolution: Developmental continuity and numeral semantics. *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)* 8. 67–75. https://doi.org/10.18653/v1/W18-0108.
Scontras, Gregory, Maria Polinsky, C.-Y. Edwin Tsai & Kenneth Mai. 2017. Cross-linguistic scope ambiguity: When two systems meet. *Glossa: A Journal of General Linguistics* 2(1). 36. https://doi.org/doi.org/10.5334/gjgl.198.
Scontras, Gregory, Michael Henry Tessler & Michael Franke. electronic. Probabilistic language understanding: An introduction to the Rational Speech Act framework. Retrieved from https://www.problang.org.
Viau, Joshua, Jeffrey Lidz & Julien Musolino. 2010. Priming of abstract logical representations in 4-year-olds. *Language Acquisition* 17(1-2). 26–50. https://doi.org/10.1080/10489221003620946.
Wu, Mien-Jen & Tania Ionin. 2019. L1-Mandarin L2-English speakers' acquisition of English universal quantifier-negation scope. *Proceedings of the Boston University Conference on Language Development* 43. 716–729.
Zaslavsky, Noga, Jennifer Hu & Roger Levy. 2021. A Rate–Distortion view of human pragmatic reasoning. *Proceedings of the Society for Computation in Linguistics* 4. 32. https://scholarworks.umass.edu/scil/vol4/iss1/32.

Zhou, Peng & Stephen Crain. 2009. Scope assignment in child language: Evidence from the acquisition of Chinese. *Lingua* 119(7). 973–988. https://doi.org/10.1016/j.lingua.2009.01.001.