Proceedings of the Society for Computation in Linguistics

Volume 4 Article 24

2021

Every ambiguity isn't syntactic in nature: Testing the Rational Speech Act model of scope ambiguity

Sherry Yong Chen
Massachusetts Institute of Technology, sychen@mit.edu

Bob van Tiel
Donders Institute for Brain, Cognition and Behaviour, bobvantiel@gmail.com

Follow this and additional works at: https://scholarworks.umass.edu/scil

 Part of the Computational Linguistics Commons, Psycholinguistics and Neurolinguistics Commons, and the Semantics and Pragmatics Commons

Recommended Citation

Chen, Sherry Yong and van Tiel, Bob (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*: Vol. 4 , Article 24.
DOI: https://doi.org/10.7275/h3rp-m711
Available at: https://scholarworks.umass.edu/scil/vol4/iss1/24

This Paper is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Proceedings of the Society for Computation in Linguistics by an authorized editor of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.

Every ambiguity isn't syntactic in nature: Testing the Rational Speech Act model of scope ambiguity*

Sherry Yong Chen
Department of Linguistics & Philosophy
Massachusetts Institute of Technology
sychen@mit.edu

Bob van Tiel
Donders Institute for
Brain, Cognition and Behaviour
bobvantiel@gmail.com

Abstract

Utterances like 'Every guest didn't leave' are ambiguous between a reading according to which no guest left and a reading according to which not all of the guests left. This ambiguity is often explained by assuming that 'every-not' utterances have two different syntactic parses. However, experimental studies have shown that pragmatic factors, such as prior probabilities and the question under discussion, also play an important role in the interpretation of ambiguous 'every-not' utterances. Recently, Scontras and Pearl (2020) put forward a probabilistic model of ambiguity resolution that makes it possible to quantify the relative contribution of syntactic and pragmatic factors. Here, we present three experiments aimed at testing this model and measuring the division of labor between syntax and pragmatics. Our results suggest that variability in the interpretation of 'every-not' utterances can be explained almost entirely in terms of pragmatics, suggesting only a marginal role for syntax.

1 Introduction

Utterances containing 'every' and 'not' may be ambiguous between two different readings:

(1) Every horse didn't jump over the fence.
   a. None of the horses jumped over the fence. ('none' reading)
   b. Not all horses jumped over the fence. ('not-all' reading)

Under the 'none' reading, none of the horses jumped over the fence. Under the 'not-all' reading, it is not the case that every horse jumped over the fence, though some may have.

This ambiguity can be explained by assuming that utterances like (1) have two possible syntactic parses. The *surface scope* parse corresponds to the surface ordering of the words in the utterance, and results in the 'none' reading given in (1a) (e.g., Carden, 1970). For the *inverse scope* parse, the negation takes scope over the universal quantifier rather than from its surface position, thus resulting in the 'not-all' reading given in (1b).

It has been observed that the two readings of (1) are not always equally accessible (e.g., Horn, 1989, p. 226–231). For example, Musolino (1998) reported that upon hearing the utterance in (1) without any context, 5-year old children generally arrived at the 'none' reading, whereas adults overwhelmingly favored the 'not-all' reading. Initially, this difference was thought to show that children only generate a subset of the syntactic parses that adults do; in particular, it was thought that children are unable to generate (or maintain in memory) a parse that involves moving constituents out of their surface positions. (Recall that the 'not-all' reading is assumed to involve the negation taking scope above the quantifier, contrary to the linear ordering.)

However, further evidence has shown that the accessibility of the two readings is modulated by pragmatic factors—both for children and adults. One such factor is the *question under discussion* (QUD) (e.g., Conroy et al., 2008; Gualmini et al., 2008). The idea underlying the notion of QUD is that discourse transpires with respect to an implicit or explicit question. Any utterance in the discourse should address this question in order to be pragmatically felicitous (Roberts, 2012). This constraint is also known as the *question-answer requirement* (Gualmini et al., 2008).

To illustrate the potential effect of the question-answer requirement on the interpretation of utterances like (1), consider the following two QUDs:

(2) a. Did any of the horses jump over the fence?

*All materials and data can be found in the GitHub repository of this project: https://github.com/linguistsherry/qud-rsa-scil2021

Proceedings of the Society for Computation in Linguistics (SCiL) 2021, pages 254-263.
Held on-line February 14-19, 2021

b. Did all of the horses jump over the fence?

The 'every-not' utterance in (1) only addresses the 'any?'-QUD in (2a) if it receives its 'none' reading. After all, on its 'not-all' reading, it could be that either none or some but not all of the horses jumped over the fence, so that the question is still unresolved. By contrast, in the case of the 'all?'-QUD in (2b), the 'none' reading provides an *overinformative* answer to the QUD, i.e., it conveys more information than a simple 'no'. Hence, it may be hypothesized that the 'all?'-QUD makes the 'not-all' reading more salient, whereas the 'any?'-QUD makes the 'none' reading more salient.

In line with this hypothesis, Gualmini and colleagues (2008) found that 5-year old children were able to access the 'not-all' reading when the 'all?'-QUD was made sufficiently salient. This finding suggests that the aforementioned difference between children and adults in the interpretation of 'every-not' utterances is more likely to have a pragmatic than syntactic source. For example, it could be that children standardly assume an 'any?'-QUD while adults are able to consider different QUDs and choose one that makes the utterance true on its QUD-appropriate reading.

Another factor that has been argued to influence the interpretation of 'every-not' utterances is listeners' *prior expectations* (Gualmini, 2004). For example, it has been found that participants are more likely to arrive at a 'not-all' reading of the ambiguous utterance in (1) if it is *a priori* likely that an 'all' situation holds, i.e., if it is likely that all of horses jumped over the fence. This finding can be explained by assuming that the probability of the 'all' situation makes the utterance of the 'every-not' utterance contextually felicitous (Wason, 1972). An alternative explanation more in line with our framework is that listeners attempt to have their interpretations differ minimally from their prior expectations (Degen et al., 2015).

In summary, the traditional idea is that the two readings of 'every-not' utterances are due to a syntactic ambiguity. However, it has since become clear that the accessibility of the readings is also heavily dependent on pragmatic factors, such as the QUD and prior expectations. It is still an open question how much syntactic and pragmatic factors impact the interpretation of 'every-not' utterances. Thus, in this paper, we investigate the division of labor between syntax and pragmatics.

To this end, we experimentally test Scontras and Pearl's (2020) computational model of ambiguity resolution which uses both syntactic and pragmatic factors as independent ingredients. In the next section, we give a brief description of the model. Afterwards, we test the model on the basis of three experiments in which we measured the salience of different QUDs (Exp. 1), listeners' prior expectations about the state of the world (Exp. 2), and their interpretation of ambiguous 'every-not' utterances such as (1) (Exp. 3). Our results suggest that the variability in the interpretation of 'every-not' utterances can be explained almost entirely on the basis of pragmatic factors, without assuming a substantial explanatory role for syntax.

The rest of the paper is organized as follows: in Section 2 we introduce Scontras and Pearl's ambiguity resolution model; Section 3 provides details of the experiments which measured human performance; Section 4 discusses results from computational modelling; Section 5 concludes.

2 Computational model

To model our data, we make use of the ambiguity resolution model formulated by Scontras and Pearl (2020). Here, we provide a concise description of the model. The interested reader is referred to Scontras and Pearl (2020) for a more detailed exposition (cf. also Savinelli et al., 2017, 2018).

The ambiguity resolution model proposed by Scontras and Pearl is an extension of the more general Rational Speech Acts (RSA) model (e.g., Frank and Goodman, 2012). The starting point of the RSA model is the stipulation of a set of possible messages M and a set of possible states S. In the case at hand, Scontras and Pearl assume there are just two possible messages: the ambiguous 'every-not' utterance and a null message, i.e.:

$$M = \{m_{\text{every-not}}, m_{\text{null}}\}$$

Conceptually, the null message can be seen as representing messages that speakers may produce whenever the ambiguous 'every-not' utterance is unsatisfactory.

Scontras and Pearl assume there are three possible states, depending on the number of individuals that satisfy the predicate (e.g., the number of horses that jumped over the fence). For convenience, we deviate notationally from Scontras and Pearl, and assume that the states indicate whether *none*, *some but not all*, or *all* of the individuals satisfy the pred-

icate, i.e.:

$$S = \{s_{none}, s_{some}, s_{all}\}$$

(Note that 'some' in s_{some} should be interpreted as 'some but not all' rather than receiving its logical interpretation of 'at least some and possibly all'.)

Next, Scontras and Pearl provide a semantics for the two messages. The null message is always true. The truth value of the ambiguous 'every-not' utterance depends on the scopal relation between the quantifier and the negation. To model syntactic scope, Scontras and Pearl assume that each message is indexed with a scope parameter i. This parameter indicates whether the message receives a surface scope parse (i = surface) or an inverse scope parse (i = inverse). The semantics of the two messages is then defined as follows (we omit the scope parameter for the null message because it is inconsequential there):

$$\begin{aligned} [\![m^{\text{surface}}_{\text{every-not}}]\!] &= \lambda s \,.\, s = s_{none} \\ [\![m^{\text{inverse}}_{\text{every-not}}]\!] &= \lambda s \,.\, s \neq s_{all} \\ [\![m_{\text{null}}]\!] &= \lambda s \,.\, s = s \end{aligned}$$

Lastly, Scontras and Pearl assume that messages are produced and interpreted relative to a question under discussion (QUD). Scontras and Pearl distinguish three QUDs, asking *whether all* individuals satisfy the predicate ($q_{\text{all?}}$), *whether any* of the individuals satisfy the predicate ($q_{\text{any?}}$), and *how many* individuals satisfy the predicate ($q_{\text{how-many?}}$), i.e.,

$$Q = \{q_{\text{all?}}, q_{\text{any?}}, q_{\text{how-many?}}\}$$

The QUD maps each state to an answer x, which can be the state itself (for $q_{\text{how-many?}}$) or a truth value (for $q_{\text{all?}}$ and $q_{\text{any?}}$). The three QUDs thus have the following semantics:

$$\begin{aligned} [\![q_{\text{all?}}]\!] &= \lambda s \,.\, s = s_{all} \\ [\![q_{\text{any?}}]\!] &= \lambda s \,.\, s \neq s_{none} \\ [\![q_{\text{how-many?}}]\!] &= \lambda s \,.\, s \end{aligned}$$

Based on the foregoing, we may define a *literal listener* L_0 who infers an answer x to the QUD q given a message m and a scope assignment i by determining, for each state, the truth value of x in that state relative to q multiplied by the truth value of the message m in that state relative to the scope assignment i, and then summing over those values:

$$P_{L_0}(x \mid m, i, q) \propto \sum_s ([\![q]\!](s) = x) \cdot [\![m^i]\!](s)$$

For example, suppose that $m = m_{\text{every-not}}$, i = surface, and $q = q_{\text{all?}}$. The probability that L_0 infers a confirmatory answer to the QUD in this case is proportional to the sum of 1 (since in s_{none} the answer to the QUD is confirmatory and the message is true), 0 (since in s_{some} the answer to the QUD is confirmatory but the message is false given the surface scope parse), and 0 (since in s_{all} the answer to the QUD is negative and the message is false), i.e., 1. In other words, L_0 necessarily infers a confirmatory answer, which is intuitively correct.

Next, Scontras and Pearl define a speaker S_1 who chooses messages with a probability that is proportional to the probability that L_0 infers the correct answer to the QUD, as follows:

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

Here, the α parameter modulates the probability that S_1 chooses the optimal message, i.e., the one for which L_0 is most likely to infer the correct answer. The probability is soft-maxed by taking the exponential.

Based on this speaker, Scontras and Pearl define a *pragmatic listener* L_1 who infers a state, scope assignment, and QUD from a message based on the probability that S_1 would produce that message given that state, scope assignment, and QUD. Importantly, the pragmatic listener also takes into consideration the prior probability of the state, scope assignment, and QUD:

$$P_{L_1}(s, i, q|m) \propto P_{S_1}(u|s, i, q) \cdot P(s) \cdot P(i) \cdot P(q)$$

Scontras and Pearl further define a *pragmatic speaker* S_2 who produces messages depending on the behaviour of L_1. However, here we are concerned with language interpretation, so we will ignore this speaker.

We wish to emphasize that, in Scontras and Pearl's model, the syntactic parse constrains but does not determine listeners' interpretation. For example, participants who arrive at an inverse scope parse will still infer the 'none' situation with a certain probability. Perhaps more surprisingly, participants who arrive at a surface scope parse will still infer the 'some but not all' situation with a certain probability insofar as the 'all?'-QUD is relevant, since the distinction between the 'none' situation and the 'some but not all' situation is irrelevant (and therefore collapses) for that QUD. Hence, it is

important to clearly distinguish the syntactic parses from listeners' interpretation.

Scontras and Pearl's model gives quantitative predictions about listeners' interpretation of ambiguous 'every-not' utterances. These predictions depend inter alia on three model-external factors: (i) the prior probability of each state $P(s)$, (ii) the probability of each scope assignment $P(i)$, and (iii) the probability of each QUD $P(q)$. We tested the model predictions by measuring participants' interpretation of ambiguous 'every-not' utterances, as well as their prior expectations about the QUD and state. Note that there is no straightforward way of measuring prior expectations about scope assignment—we explore different possibilities in the 'Model evaluation' section.

3 Experiments

To test how adult speakers interpret utterances such as (1), we constructed 72 short stories while separately manipulating the prior probability and biases toward certain types of QUDs. A set of sample stimuli can be found in Table 1.

The stories were constructed as follows. First, we constructed 24 stories that intuitively varied in which state of the world was most likely: 8 stories made the 'none' situation likely, 8 stories made the 'some' situation likely, and 8 stories made the 'all' situation likely. Second, we expanded each story with one of three QUD-biasing sentences: 1 sentence made the 'any?'-QUD likely, 1 sentence made the 'how-many?'-QUD likely, and 1 sentence made the 'all?'-QUD likely. Third, we expanded each story with an ambiguous 'every-not' utterance made by one of the characters. These utterances always had the negation contracted to the verb (e.g., 'hasn't' rather than 'has not').

We conducted 3 experiments: Exp. 1 tested participants' prior expectations about the state of the world for each story, Exp. 2 tested participants' intuitions about the salience of different QUDs for each story. The results of these two experiments are used to parametrize the models. Exp. 3 probed participants' interpretation of 'every-not' utterances. All experiments were held on IbexFarm (Drummond, 2013), and each experiment has 45 participants recruited via Prolific. Statistical analysis and modelling was carried out using R (R Core Team, 2017)

3.1 Experiment 1: Situation priors

Exp. 1 measured the prior probabilities of the 'none', 'some but not all', and 'all' situations. 45 participants were presented with 24 short stories, in which we manipulated whether the 'none', 'some but not all', or 'all' situation was intuitively more likely. The stories were presented without the sentence that made one of the QUDs salient, and without the ambiguous 'every-not' utterance. Participants had to rate the prior probability of each situation by moving continuous sliders. The complete experiment can be accessed at http://spellout.net/ibexexps/qud-project/prior_list1/experiment.html.

Fig. 1 shows the ratings for the prior probabilities of the three situations, depending on our intuitive classification. It can be seen that participants' ratings generally followed our intuitions. However, participants had an overall strong preference for the 'some but not all' situation, perhaps because that was the least extreme option. Nonetheless, it is clear that the ratings for the 'none' situation were the highest for the stories that we intuitively classified as making that situation most likely, and the same holds, mutatis mutandis, for the ratings for the 'some but not all' and 'all' situations.

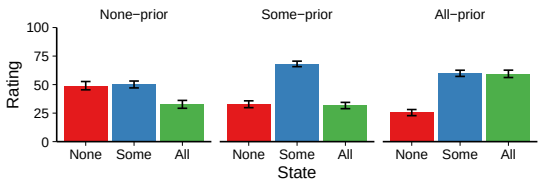


Figure 1: Raw ratings for the prior probability of each state (Exp. 1). The different facets refer to our intuitive classification.

3.2 Experiment 2: QUD salience

Exp. 2 measured the salience of the 'any?', 'all?', and 'how-many?' QUDs. 45 participants were presented with 24 (out of 72 in total) short stories, which made one of the three QUDs intuitively salient. The story selection was varied across 3 lists. The stories were presented without the ambiguous 'every-not' utterance. Participants had to rate the salience of each QUD by moving continuous sliders. The complete experiment can be accessed at http://spellout.net/ibexexps/qud-project/qud_list1/experiment.html.

Fig. 2 shows the ratings for the salience of the three QUDs, depending on our intuitive classifica-

	All-prior	Some-prior	None-prior
Context	Rachel is an industrious teaching assistant who has to grade student essays. After a week, her boss asks...	Barney has been submitting recipes to a famous cooking website for some time now. His partner asked...	Jack is an incompetent homicide detective trying to solve three recent murders. His boss tells him that...
All?-QUD	if she is ready to enter the grades into the school system.	whether all of his submissions had been posted on the website.	he will have to look for a new job if he doesn't solve these murders.
HowMany? QUD	how much progress she's made so far.	how successful his submissions had been.	he will get a bonus depending on his performance.
Any? QUD	if she has already started grading the essays.	whether any of his submissions had even been posted on the website.	the newspapers will be all over him if he fails to solve any of these murders.
'Every-not' utterance	Rachel says: Every essay hasn't been graded.	Barney says: Every submission hasn't been published.	Jack says: Every case hasn't been solved.

Table 1: Sample stimuli from Exps. 1–3

tion. It can be seen that participants' ratings followed our intuitions, i.e., the 'any?'-QUD was most salient for the stories that we intuitively thought would make that QUD most salient, and the same holds, mutatis mutandis, for the other two QUDs.

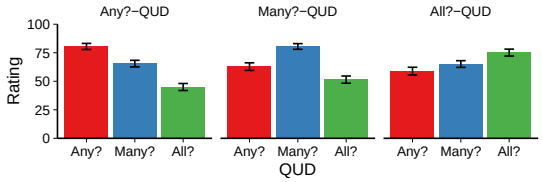


Figure 2: Raw ratings for the salience of each QUD (Exp. 2). The different facets refer to our intuitive classification.

However, the QUD was also influenced by prior expectations about the state of the world, as shown in Fig. 3. That is, participants were more likely to rate the 'any?'-QUD as salient if the 'none' situation was *a priori* more likely. Similarly, participants were more likely to rate the 'all?'-QUD as salient if the 'all' situation was *a priori* more likely. This interdependence makes intuitively sense: when certain situations are very likely, any deviations from those situations thereby become highly relevant, and thus a potential QUD.

However, it is important to note that the notions of prior probabilities and QUD salience were not completely equivalent. To illustrate, Fig. 4 shows the salience ratings of each QUD plotted for each of our intuitive classifications about both QUD salience and prior expectations. This figure shows, e.g., that when the 'any?'-QUD was salient according to our intuitions, the 'any?'-QUD was indeed

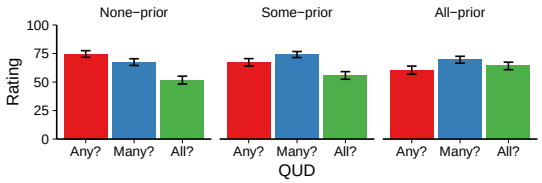


Figure 3: Raw ratings for the salience of each QUD (Exp. 2). The different facets refer to our intuitive classification about prior expectations.

rated highest for all prior expectation conditions, though this effect was most prominent when the prior expectations also favored a 'none' state.

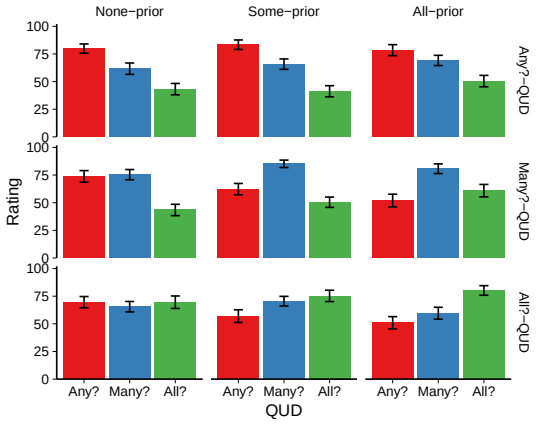


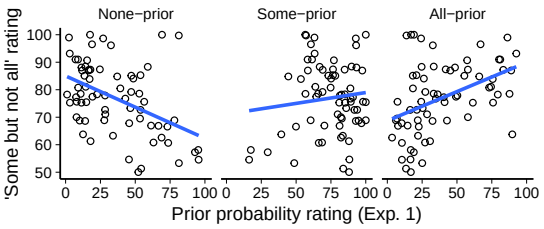
Figure 4: Raw ratings for the salience of each QUD (Exp. 2). The different facets refer to our intuitive classification about prior expectations and QUD salience.

3.3 Experiment 3: Resolving scope ambiguity

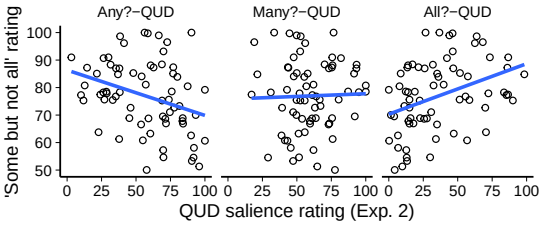
Exp. 3 tested participants' interpretation of ambiguous 'every-not' utterances. 45 participants were presented with 24 (out of 72) short stories. The story selection was varied across 3 lists. Each story ended with one of the characters producing an 'every-not' utterance. Participants had to indicate their interpretation of this utterance by rating the probability of the 'none', 'some but not all', and 'all' situations, given the speaker's utterance. The experiment can be accessed at http://spellout.net/ibexexps/qud-project/scope_list1/experiment.html.

Overall, in line with prior research, participants gave higher ratings to the 'some but not all' situation than to the 'none' situation (77 vs. 39). Indeed, for only 5 (of the 72) items did participants on average assign a higher rating to the 'none' situation than to the 'some but not all' situation. This observation already foreshadows the conclusion that participants had an overwhelming preference for an inverse scope parse of the target utterance. Further in line with expectations, the 'all' situation, in which the 'every-not' utterance was false on both of its readings, only received a very low rating (5). Fig. 5 shows the ratings for the 'some but not all' situation plotted alongside the prior probability ratings from Exp. 1 and the QUD salience ratings from Exp. 2. The ratings for the 'some but not all' situation are interesting because they conclusively show that participants arrived at a 'not-all' interpretation of the ambiguous 'every-not' utterance.

Fig. 5 shows that ratings for the 'some but not all' situation (i.e., endorsements of the 'not-all' reading) increase when the 'all' situation is a priori more likely and, less forcefully, when the 'some' situation is a priori more likely. Conversely, ratings for the 'some but not all' situation decrease when the 'none' situation was rated as a priori more likely. In line with prior research, participants gave higher ratings to the 'some but not all' situation when the 'all?'-QUD was more salient, and lower ratings when the 'any?'-QUD was more salient. There was no noticeable effect of the salience of the 'how-many?'-QUD on the ratings for the 'some but not all' situation.



(a) Correlation between the rating for the 'some but not all' situation (Exp. 3) and the ratings for prior probabilities of the three situations (Exp. 1).



(b) Correlation between the rating for the 'some but not all' situation (Exp. 3) and the ratings for QUD salience of the three possible QUDs (Exp. 2).

Figure 5: Correlation between the rating for the 'some but not all' situation (Exp. 3) and the results of Exps. 1 and 2.

4 Model evaluation

The goal of this paper is to evaluate how accurately Scontras and Pearl's ambiguity resolution model captures the way participants in Exp. 3 interpreted 'every-not' utterances. However, we also want to measure the relative contribution of each of the components of the model, i.e., prior expectations about the state of the world, the question under discussion (QUD), and syntactic scope. With the latter goal in mind, we tested both the full ambiguity resolution model L_{full} (i.e., Scontras and Pearl's pragmatic listener L_1) and four models that systematically nullified the effect of one of the components, as follows:

- $L_{noPrior}$ did not take into account prior probabilities over possible states (Exp. 1).
- L_{noQUD} did not take into account the QUD (Exp. 2).
- $L_{surface}$ always assigned surface scope.
- $L_{inverse}$ always assigned inverse scope.

The ambiguity resolution model associates 'every-not' utterances with (inter alia) a probability distribution over possible states of the world, depending on prior expectations about the state of the world, the QUD, and the syntactic scope. To obtain predictions from the model, we parametrized these components on the basis of the results of Exp. 1 (prior expectations about the state of the world) and

Exp. 2 (salience of the QUD). In line with Scontras and Pearl, we assume that the models consider both syntactic scopes equiprobable, except of course for the $L_{surface}$ and $L_{inverse}$ models which always converged on one of both scope options.

Thus, we generated, for each short story and for each model, a probability distribution over possible states of the world (i.e., s_{none}, s_{some}, and s_{all}). To obtain a quantitative measure of model fit, we calculated, for each model, three correlations. First, we calculated the mean correlation between predictions and normalized data for each item. Second, we calculated the mean correlation between predictions and normalized data for each possible state of the world. Third, we calculated the overall correlation between predictions and normalized data. The three correlations are shown in Table 2. Fig. 6 visualizes the overall correlation for each model.

The correlations between predictions and data for the full model are generally quite high. The only exception is the per state correlation, which lies slightly below 0.4. Note, however, that this correlation was skewed down because the model made relatively poor predictions about the variability in the ratings for the 'all' situation, in which the 'every-not' utterance was false. (Note that the model assigns some probability to that state in case the 'any?'-QUD is highly salient.)

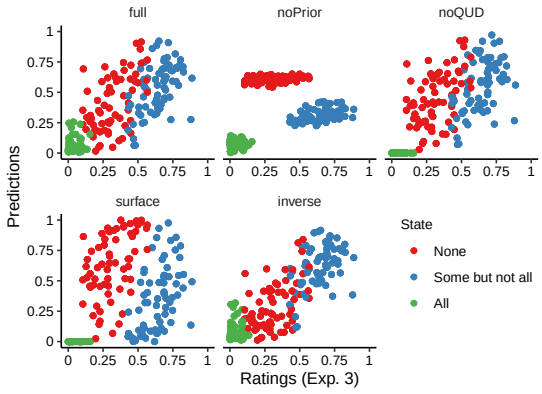


Figure 6: Scatterplot of predictions and data for each item, state of the world, and ambiguity resolution model.

The results indicate that prior expectations about the state of the world are highly relevant, as the correlation drops from .67 to .34 once these are removed from the model. By contrast, the effect of QUD is only marginal: removing this factor from the model reduces the correlation from .67 to .65. A model that always assigned a surface scope parse performed poorly in comparison to the full model in which both scope options were deemed equiprobable. Interestingly, however, the model that always assigned an inverse scope parse outperformed all other ambiguity resolution models—including the full model.

To confirm whether the optimal model was indeed one in which the 'every-not' utterance unambiguously received an inverse scope parse, we formulated a model with scope preference as a free parameter, and subsequently optimized this model. The result of this analysis confirmed that the optimal model was one with an invariant preference for inverse scope. In other words, the optimal model was one that completely ignored the possibility of a surface scope parse of the sentence.

Scontras and Pearl conclude that "pragmatic factors play a larger role than grammatical processing factors" (p. 1) in explaining the interpretation of 'every-not' utterances. Our results suggest an even bolder claim: the way people interpret 'every-not' utterances is exhaustively shaped by pragmatic factors (i.e., the QUD and, especially, prior expectations about the state of the world) rather than by their vaccillation between different syntactic parses. More specifically, our results suggest that English language users consistently assign an inverse scope parse to 'every-not' utterance.

5 General discussion

In this paper, we experimentally investigated listeners' interpretation of ambiguous 'every-not' utterances, such as 'Every horse didn't jump over the fence'. Such utterances are ambiguous between a 'none' reading ('None of the horses jumped over the fence') and a 'not-all' reading ('Not all horses jumped over the fence'). In particular, we studied how listeners' interpretation of such utterances is influenced by their prior expectations about the state of the world and by the question under discussion (QUD).

To computationally model the data, we made use of the ambiguity resolution model described by Scontras and Pearl (2020) (cf. also Savinelli et al., 2017, 2018). This model assumes that listeners' interpretations of ambiguous utterances are shaped by (i) their syntactic parse, (ii) their prior expectations about the state of the world, and (iii) the QUD. We parametrized the factors in (ii) and (iii) based on experimental data. Scontras and Pearl's model generally offered a fair approximation of the actual

	L_{full}	$L_{noPrior}$	L_{noQUD}	$L_{surface}$	$L_{inverse}$
per item	.82	.42	.81	.54	.88
per state	.39	.22	.34	.31	.36
overall	.78	.39	.79	.48	.85
average	.67	.34	.65	.45	.70

Table 2: Correlations between data from Exp. 3 and predictions for each ambiguity resolution model. The optimal correlation is marked in green.

data. Interestingly, model comparison indicated that participants' interpretation was mostly shaped by their prior expectations, and only marginally by the QUD, which goes against remarks made in the literature (e.g., Gualmini et al., 2008).

However, we also observed an interaction between prior expectations and QUD salience. Such an interaction makes intuitive sense. For example, if it is extremely unlikely that any of the horses jumped over the fence, the QUD 'Did any of the horses jump over the fence?' thereby becomes more salient. Conversely, if someone asks 'Did all of the horses jump over the fence?', a state of the world in which all of the horses jumped over the fence thereby becomes more likely, since it is entertained as a real possibility by a presumably rational questioner. It is possible, then, that prior evidence for the effects of QUD may have been confounded by effects of prior expectations about the state of the world. Scontras and Pearl's model assumes that prior expectations and the QUD are completely independent: our results call that assumption into question.

A second main finding from the model evaluation is that the data was best described by a model that did not assume that 'every-not' utterances are ambiguous at the syntactic level. The optimal model always assigned an inverse scope parse, according to which the negation was moved out of its surface position to take scope over the quantifier. The data was thus best described by a model that only made use of pragmatic cues.

There are two ways of interpreting this finding. The first is by concluding that 'every-not' utterances receive an inverse scope parse by default, and that their apparent ambiguity is wholly determined by pragmatic factors. However, it could also be that there is an interdependence between syntactic scope and the pragmatic factors that we tested. For example, it could be that prior expectations about the state of the world shape listeners' interpretation indirectly by making one syntactic parse more available. The computational model that we tested does not factor in such potential interactions between syntax and pragmatics, as it includes these ingredients as completely independent factors. We leave for future work the suggestion of developing a model that also takes into consideration possible interdependencies between the various pragmatic and syntactic cues.

In any case, we do not necessarily want to argue that 'every-not' utterances lack any syntactic ambiguity. Indeed, ongoing research by Ira Noveck and Kazuko Yatsushiro suggests substantial cross-linguistic differences in the accessibility of the two readings for 'every-not' utterances. Such findings are reminiscent of earlier disputed arguments that the interpretation of 'every-not' utterances in English systematically varies across dialects (e.g., Carden, 1970; Baltin, 1974; Horn, 1989). Intuitively, such differences are more likely to be structural than pragmatic in nature. For example, it seems unlikely that Dutch participants tended to imagine different QUDs than French participants. Therefore, here we merely claim that participants' interpretation of these sentences can be modelled largely in terms of pragmatics rather than syntax.

What complicates matters is that the model has a number of free parameters that have important effects on the model comparisons. One such free parameter is the rationality parameter α that is standardly set to 1. Another free parameter is the set of alternatives. We followed Scontras and Pearl in assuming that the only alternative to the 'every-not' utterance is a null message that is always true. Scontras and Pearl's motivation for this decision is that participants in experiments often only see 'every-not' utterances. Indeed, this was also the case in our experiment. Still, it seems that the 'every-not' utterance itself can already make certain alternatives salient. For example, it is often assumed that an utterance like 'Every horse didn't jump over the fence' competes with the simpler 'No horse jumped over the fence' (Horn, 1989, p. 226ff.). Indeed, various authors have argued that the relative accessibility of the inverse scope parse compared to the surface scope parse (at least in English) is due to the 'no' utterance offering a simpler

way of conveying the reading corresponding to a surface scope parse. Similarly, a classic observation is that negated sentences generally make their positive counterparts salient (e.g., Wason, 1972). Exploratory analyses where we included 'no' utterances and 'every' utterances as alternatives suggest a larger role for syntax compared to the analyses presented in this paper. However, before drawing any firm conclusions, one will need to find a way to determine which alternatives (if any) participants take into consideration when they interpret 'every-not' utterances.

The prominence of pragmatic factors calls for a re-evaluation of some of the key findings from the literature. Perhaps the most widely discussed observation is that in null contexts children tend to arrive at a 'none' reading whereas adults prefer the 'not-all' reading (Musolino, 1998; Gualmini et al., 2008). Initially, this difference was taken to indicate that children are only able to access the 'none' reading. However, Gualmini and colleagues (2008) later showed that children can in fact arrive at a 'not-all' reading if the QUD targets whether or not, e.g., all of the horses jumped over the fence. This finding has been interpreted as evidence that children may access an inverse scope parse. However, Scontras and Pearl's computational model—validated by our data—shows that a 'not-all' reading may in fact emerge from a surface scope parse if certain pragmatic constraints are satisfied, e.g., if the QUD asks whether all of the horses jumped over the fence and if a situation where none of the horses jumped over the fence is unlikely. Hence, it could be that the difference between children and adults is more reflective of a difference in how they contextualize the target utterances.

Similarly, Syrett and colleagues (2014) observed that prosody can steer listeners to either a 'none' or 'not-all' reading. They interpreted this finding as showing that prosody can steer listeners' syntactic parsing to either a surface scope or inverse scope parse. As Syrett and colleagues acknowledge, however, it is well known that prosody also helps listeners to select the most plausible QUD (e.g., Kadmon and Roberts, 1986). While Syrett and colleagues controlled for prior probabilities, some of their results may be explained based on the assumption that the prosodic manipulations (to the extent that they were effective) determined the most likely QUD rather than which syntactic parse to choose—assuming, as before, that these factors work largely independently. This alternative explanation might also account for the observation that prosody only has a relatively mild effect on how people disambiguated utterances with a quantificational ambiguity.

The more general conclusion from our study is that many of the experimental findings on 'every-not' utterances may be amenable to a purely pragmatic explanation, without having to assume that listeners routinely alternate between syntactic parses. In order to evaluate this hypothesis, it is necessary to carry out experiments that control for pragmatic factors, and to formally evaluate whether it is necessary to postulate differences in the accessibility of the syntactic parses. Such an ambitious enterprise must be left to future research.

6 Acknowledgments

This research was funded by the German Research Council (DFG FR 3482/2-1, KR951/14-1, SA 925/17-1) within SPP 1727 (XPrag.de), and by the Dutch Science Organisation (Gravitation grant 'Language in Interaction', 024.001.006). We thank the reviewers for their thorough and valuable feedback, which have led us to make important improvements on the contextualization of our research and the interpretation of our results.

References

Mark Baltin. 1974. Quantifier-negative interaction. In *New ways of analyzing variation in English*, pages 30–36, Washington, DC. Georgetown University Press.

Guy Carden. 1970. A note on conflicting idiolects. *Linguistic Inquiry*, 1:281–290.

Anastasia Conroy, Scott Fults, Julien Musolino, and Jeffrey Lidz. 2008. Surface scope as a default: The effect of time in resolving quantifier scope ambiguity. Poster presented at the 21st CUNY conference.

Judith Degen, Michael Henry Tessler, and Noah D Goodman. 2015. Wonky worlds: Listeners revise world knowledge when utterances are odd. *Proceedings of the 37th Annual Conference of the Cognitive Science Society*, pages 548–553.

Alex Drummond. 2013. Ibex Farm. https://spellout.net/ibexfarm/.

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

Andrea Gualmini. 2004. *The ups and downs of child language*. New York, NY: Routledge.

Andrea Gualmini, Sarah Hulsey, Valentine Hacquard, and Danny Fox. 2008. The question–answer requirement for scope assignment. *Natural Language Semantics*, 16:205.

Laurence R. Horn. 1989. *A Natural History of Negation*. CSLI, Stanford, CA.

Nirit Kadmon and C. Roberts. 1986. Prosody and scope: The role of discourse structure. In *Proceedings of the Annual Meeting of Chicago Linguistics Society*, pages 16–28. Chicago Linguistic Society.

Julien Musolino. 1998. *Universal grammar and the acquisition of semantic knowledge*. Ph.D. thesis, University of Maryland.

R Core Team. 2017. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics & Pragmatics*, 5:1–69.

KJ Savinelli, Gregory Scontras, and Lisa Pearl. 2017. Modeling scope ambiguity resolution as pragmatic inference: Formalizing differences in child and adult behavior. In *Proceedings of the Annual Meeting of Cognitive Science Society*.

KJ Savinelli, Gregory Scontras, and Lisa Pearl. 2018. Exactly two things to learn from modeling scope ambiguity resolution: Developmental continuity and numeral semantics. In *Proceedings of the 8th Workshop on CMCL*, pages 67–75.

Gregory Scontras and Lisa Pearl. 2020. When pragmatics matters more for truth-value judgments: An investigation of quantifier scope ambiguity. Lingbuzz/005287.

Kristen Syrett, Georgia Simon, and Kirsten Nisula. 2014. Prosodic disambiguation of scopally ambiguous quantificational sentences in a discourse context. *Journal of Linguistics*, 50:453–493.

Peter Wason. 1972. The context of plausible denial. *Journal of Verbal Learning and Language Behavior*, 4:7–11.