When Parsing and interpretation misalign: a case of wh-scope ambiguity in Mandarin

Ming Xiang
Department of Linguistics
The University of Chicago
Chicago, IL, 60637, USA
mxiang@uchicago.edu

Zhewei Dai
Department of Mathematics and Computer Science
Alma College
Alma, MI, 48801, USA
dai@alma.edu

Suiping Wang
Philosophy and Social Science Laboratory
of Reading and Development in Children and Adolescents
Ministry of Education
South China Normal University
Guangzhou, 510631, China
wangsuiping@m.scnu.edu.cn

When Parsing and interpretation misalign: a case of wh-scope ambiguity in Mandarin

Abstract

A great amount of sentence processing work has focused on revealing how the parser incrementally integrates each incoming word into the current linguistic representation. It is often explicitly or implicitly assumed that the structure endorsed by the parser would determine the ultimate interpretation of the sentence. The current study investigates whether the interpretive bias in sentence comprehension necessarily tracks the parsing bias. Our case study concerns the locality bias in non-local dependencies, specifically, the Mandarin wh-in-situ scope dependencies. Our findings suggest a misalignment between parsing and interpretative decisions at the global level. In particular, for Mandarin wh-in-situ constructions that involve scope ambiguity, there is a locality bias in parsing, but there is an anti-locality bias in interpretation. Building upon the Rational Speech Act framework, We propose a bayesian pragmatic analysis to account for these findings. Under our proposal, the seeming conflict between parsing and interpretation will ultimately disappear because parsing preferences will be naturally embedded under the pragmatic reasoning process to generate the ultimate interpretation. The current study therefore makes novel contributions, both empirically and theoretically, to address the broader question about the relationship between parsing and interpretation.[*]

Keywords: parsing, sentence comprehension, bayesian pragmatic reasoning, long distance dependencies, locality effect, Chinese wh-in-situ

[*]This material is based on work supported by the National Science Foundation under Grant No. BCS1451635 and by the University of Chicago Humanities Division Council. We would like to thank Whitney Tabor, two anonymous reviewers and editors of *Language* John Beavers and Ezra Keshet for their constructive feedback. Their comments greatly improved this work. All remaining errors are due to the authors.

1. Introduction. Sentence comprehension requires a parser that establishes the structural representation of the to-be-interpreted sentence. A great amount of sentence processing work has focused on revealing how the parser incrementally integrates each incoming word into the current grammatical representation. As for the mapping between structure and interpretation, it is often explicitly or implicitly assumed that the structure endorsed by the parser should determine the ultimate interpretation of the sentence. This seemingly simple mapping between parsing and interpretation, however, faces challenges from observations showing that the interpretations comprehenders obtain could sometimes deviate from what the grammatical parse allows. A salient case of this came from sentences containing temporary garden-path ambiguities (e.g. Christianson et al. 2001, Qian et al. 2018). For example, Christianson and colleagues (2001) examined people's interpretation for the temporarily ambiguous but globally unambiguous sentence "While Anna dressed the baby that was cute and cuddly played in the crib". When participants were asked "Did Anna dress the baby", the majority of the responses were "yes", even though this interpretation is incompatible with the grammatical parse at the global level. Misinterpretations also arise for "local coherence" sentences such as "The coach smiled at the player tossed the frisbee" (Tabor et al. 2004, Konieczny et al. 2009), resulting in an interpretation that the player tossed the frisbee when the grammatically licensed interpretation is someone else tossed the frisbee to the player. It has also been found that for relatively infrequent constructions or sentences with non-canonical word orders, such as passives or object-clefts, misinterpretation may happen when the grammatically licensed interpretation is implausible (Ferreira 2003). Different proposals have been put forward to deal with the attested mismatches between parsing and interpretation. For example, the *good enough* approach to comprehension (Ferreira et al. 2001, 2002, Christianson et al. 2001, Ferreira & Patson 2007) explains such findings by allowing interpretations to be derived through simple heuristics (e.g. world knowledge, word order, etc) rather than fully specified structural parses. The *noisy channel* account (Levy 2008, Gibson et al. 2013), on the other hand, accounts for the empirical findings by introducing noise or uncertainty on the linguistic input a comprehender perceives.

The current study has two goals, one empirical and the other theoretical. First, we identify a new empirical case unrelated to the previous observations discussed above, that also demonstrates (descriptively speaking) a misalignment between parsing and interpretation. Our case study concerns a particular kind of structural ambiguity, the wh-scope ambiguity in Mandarin wh-in-situ construction. There is a large body of previous work on structural ambiguity resolution (e.g. Frazier & Fodor 1978, MacDonald et al. 1994, Cuetos & Mitchell 1988, Van Gompel et al. 2000). The primary research

question in this literature is to understand the parsing biases of structurally ambiguous sentences and also explain the sources for such parsing biases. The primary question of the current study is different. Our empirical focus is not only the parsing bias itself, but more importantly the misalignment between the parsing bias and the interpretation bias. Theoretically, our proposal for the current empirical findings offers a new kind of analytical possibility to address the more general question about the relationship between parsing and interpretation. Specifically, we will propose a Bayesian account that integrates parsing information with pragmatic reasoning to predict quantitative results on interpretation. Our Bayesian account is built upon the Rational Speech Act framework (RSA, Goodman & Frank 2016). Under our account, the seeming mismatch between parsing and interpretation will ultimately disappear, since parsing preferences will be integrated into a general pragmatic reasoning process to derive the ultimate interpretation.

Our study examines the locality effect in wh-in-situ constructions that show scope ambiguity. More details about the wh-in-situ constructions will be introduced in the next section, but generally speaking, locality bias is commonly observed in sentence parsing. A representative example of this is the well-documented distance effect in processing non-local dependencies. In constructions that involve non-local dependencies, such as in English relative clauses or wh-questions, it is often observed that greater distance between the two elements on a dependency chain enhances processing difficulty, as measured by decreased acceptability judgments, increased reading time or enhanced neurophysiological responses (Gibson 1998, Warren & Gibson 2002, Van Dyke & Lewis 2003, Lewis & Vasishth 2005). As an example, consider 1 from Alexopoulou and Keller (2007). In their results, as the distance between the verb *fire* and its fronted wh-argument *who* increased from 1a to 1c, the acceptability rating decreased accordingly.

(1) a. Who will we fire?
    b. Who does Mary claim we will fire?
    c. Who does Jane think Mary claims we will fire?

The shorter dependency is generally more preferred to the longer ones, hence the *locality* bias. Many accounts of this effect are based on hypotheses about how working memory is structured and deployed to support language comprehension. For example, in Dependency Locality Theory (Gibson 1998, 2000), the processing cost for completing a dependency is a function of the number of discourse referents between the two elements on a dependency chain. Under the memory retrieval account (Lewis & Vasishth 2005), processing cost is in large part determined by how quickly and unambiguously the relevant dependent element can be retrieved from working memory, amongst all other memory

representations that could potentially introduce interference. Longer dependencies are more likely to introduce elements that can interfere with the retrieval target, leading to an increased processing cost.

Taking advantage of the well-established parsing preference for shorter dependencies, the current study examines people's interpretation of scope-ambiguous sentences in the presence of clear locality parsing bias. In section 2, we establish the empirical generalization that even though the shorter scope dependency is the preferred structural parse over the longer one, consistent with the broader conclusion about locality bias in parsing; the interpretation obtained by the comprehenders nonetheless aligns with the longer scope dependency. To account for this, we develop a proposal in section 3 and 4 that integrates parsing biases with Bayesian pragmatic reasoning. We discuss the implications and remaining questions of the current proposal in section 5.

2. Parsing and interpreting wh-in-situ scope – locality and anti-locality.

2.1. Locality bias in parsing. In Mandarin Chinese, a wh-in-situ language, a covert dependency is formed between an in-situ wh-phrase and its scope position (Aoun & Li 1993, Cheng 1991, 2003, Huang 1982, Tsai 1994). An example of a Mandarin Chinese wh-construction is given in 2:

(2)

记者们	知道 [$_{Clause1}$	市长	严惩了	哪些	官员。]
jizhemen	zhidao	shizhang	yancheng-le	naxie	guanyuan
Reporter	know	mayor	punish-perf	which	official

"The reporters knew which officials the mayor punished."

The example in 2 presents an embedded wh-question: the wh-element *which official* takes scope over the embedded clause. Despite the lack of overt cues that signal a non-local dependency, processing evidence from Xiang and colleagues (2015, 2020) showed that the incremental construction of a wh-in-situ dependency is constrained by the same parsing principles that regulate the processing of overt non-local dependencies.

More important for the current purpose, based on experimental evidence, Xiang and Wang (2020) argued that when there is scope ambiguity for a wh-in-situ element, the local scope dependency (low scope) is less costly than the non-local high scope dependency, essentially illustrating a locality bias like their overt-dependency kin in English. This conclusion is largely based on a comparison between two types of sentences, as shown by the examples in 3:

(3) a. 记者们 知道 [$_{Clause1}$市长 透露了 [$_{Clause2}$市政府 严惩了

jizhemen zhidao shizhang toulu-le shizhengfu yancheng-le
Reporter know mayor reveal-perf city-council punish-perf
哪些-官员。]]
naxie-guanyuan
which-CL-official
"The reporters knew which officials the mayor revealed that the city council punished." (High Scope) OR
"The reporters knew the mayor revealed which officials that the city council punished." (Low Scope)

b. 记者们 知道 [$_{Clause1}$市长 相信 [$_{Clause2}$市政府 严惩了
jizhemen zhidao shizhang xiangxin shizhengfu yancheng-le
Reporter know mayor believe city-council punish-perf
哪些-官员。]]
naxie-guanyuan
which-CL-official
"The reporters knew which officials the mayor believed that the city council punished." (High Scope)
(unavailable): "The reporters knew the mayor believed which officials the city council punished." (Low Scope, blocked)

The sentence in 3a is ambiguous since the wh-in-situ item could take either high scope over clause 1 or low scope over clause 2. The low scope, i.e. the local scope dependency that associates the wh-item with a scope position at the left edge of clause 2, was argued by Xiang and colleagues to be preferred over the high scope. The critical argument for this conclusion comes from the comparison between 3a and 3b. The two sentences in 3a and 3b are almost identical, except that the lower verb *believe* in 3b is lexically constrained such that it does not allow an embedded interrogative clause as its complement. Such a sub-categorization constraint on verbs is well-known in the literature (Ginzburg 1995), and we give some examples of verbs with distinct sub-categorization properties in 4. Verbs like *know* or *reveal* allow either embedded interrogative or declarative complement clauses, as shown in 4a and 4b. But verbs like *believe* or *think* only allow embedded declaratives, as shown by the contrast in 4c and 4d.

(4) a. John *knew/revealed* who wrote that book.
b. John *knew/revealed* Mary wrote that book.
c. * John *believed/thought* who wrote that book.
d. John *believed/thought* Mary wrote that book.

Given the verb difference between 3a and 3b, one important consequence is that the low-scope dependency in 3b is blocked. Xiang and Wang (2020) argued that sentences like 3b show substantial parsing difficulty because the low-scope dependency is expected but blocked, resulting in a much lower acceptability rating for 3b than 3a. One may ask whether the low acceptability for the high-scope only 3b is indeed due to the unavailability of the low-scope dependency, rather than the fact that 3a is scope-ambiguous and it could have benefited from an ambiguity-advantage effect (e.g. Traxler et al. 1998). The critical observation that argues against the ambiguity advantage explanation is that if one switches the position of the verbs *know* and *reveal* in 3a, as well as the position of the verbs *know* and *believe* in 3b, the previously observed acceptability differences between the two conditions disappears. The two relevant conditions are shown in 5:

(5) a. 记者们 透露了 [$_{Clause1}$市长 知道 [$_{Clause2}$市政府 严惩了
jizhemen toulu-le shizhang zhidao shizhengfu yancheng-le
Reporter reveal-perf mayor know city-council punish
哪些-官员。]]
naxie-guanyuan
which-CL-official
"The reporters revealed which officials the mayor knew that the city council punished." (High Scope) OR
"The reporters revealed the mayor knew which officials that the city council punished." (Low Scope)

b. 记者们 相信 [$_{Clause1}$市长 知道 [$_{Clause2}$市政府 严惩了
jizhemen xiangxin shizhang zhidao shizhengfu yancheng-le
Reporter believe mayor know city-council punish
哪些-官员。]]
naxie-guanyuan
which-CL-official
(unavailable): "The reporters believed which officials the mayor knew that the city council punished." (High Scope, blocked)
"The reporters believed the mayor knew which officials the city council punished." (Low Scope)

Parallel to 3a and 3b, 5a is scope ambiguous and 5b is not. But the unambiguous 5b, critically different from the unambiguous 3b, only has the low-scope parse, since the high scope is blocked by the matrix verb 'believe'. There is no acceptability difference between 5a and 5b, in contrast to the acceptability difference between 3a and 3b.[1] This

contrast lends strong support to the conclusion that there is a locality bias in parsing. Whenever a local dependency is available, it is relatively easy for the parser to successfully establish a syntactic parse, as in the case of 3a, 5a and 5b; but when the local dependency is blocked, as in 3b, the parser encounters a greater degree of parsing difficulty. The alternative account based on the ambiguity advantage effect, on the other hand, would make the wrong prediction that the ambiguous 5a should be rated much higher than the unambiguous 5b.

Building upon the observation that when there is scope ambiguity for wh-in-situ expressions, there is a strong preference for the local scope parse, the main empirical question of the current study is to identify the interpretation bias people have for scope ambiguous wh-constructions. To start with, if interpretation bias tracks parsing bias, one reasonable hypothesis is that the scope ambiguity should ultimately be resolved to favor interpretations supported by the local scope dependency. We test this possibility in Experiment 1 using a truth value judgment task.

2.2. Experiment 1: Scope interpretation bias – A truth value judgment task. In this experiment, participants were presented with a sentence containing a wh-in-situ expression. The target sentence by itself can have different interpretations, depending on whether it is parsed as having a low scope (local) dependency or a high scope dependency. Prior to the target sentence, the participants were also presented with a context scenario that was only compatible with the interpretation of one of the parses. They were instructed to judge whether the target sentence *fits* the context. Their judgments, therefore, can provide us with some evidence as to which scope dependency they have committed to. Consider a target sentence like the one in 6:

(6) 艾米丽 公布了 [$_{Clause1}$她的团队 发现了 [$_{Clause2}$外星人 建造了
Emily gongbu-le tade tuandui faxian-le waixingren jianzao-le
Emily announce-perf her team discover-perf aliens establish-perf
哪座城市。]]
na-zuo chengshi
which-CL-city
High scope: "Emily announced which city her team discovered aliens established."
OR
Low scope: "Emily announced her team discovered which city the aliens established."

When the high scope reading is true, the sentence can be roughly paraphrased as "Emily announced the answer to the question, 'which city did Emily's team discover the aliens

established?' ". This reading entails that Emily revealed the identity of the city. Suppose the answer to the embedded question is "Rome", then the high scope reading means that Emily revealed that her team discovered that the aliens established Rome. The low scope reading, on the other hand, can be paraphrased as "Emily announced her team discovered the answer to the question 'Which city did the aliens establish?' ". This reading, crucially, does not necessarily entail Emily revealed the identity of the city. This interpretation difference between the high and low scope dependencies will play an important role in our experiment below.

Material, participants and procedure. We constructed four different conditions. An example is shown in 7. These conditions share the same context scenario, and differ from each other in the target sentences. Participants were instructed to judge whether the meaning of the target sentence *match* or *does not match* the context scenario. For convenience, we will refer to the task as a truth value judgment task, and code the *match* and the *does not match* responses as *true* and *false* judgments respectively. The first condition 7a has a target sentence like 6. The preceding context makes the high-scope construal *true* and the low-scope construal *false* for the target sentence in 7a. To counterbalance the association between the True/False judgments and the high/low scope construals of the target sentences, we modified the matrix predicate in 7a to create the condition in 7b. In 7b, the matrix verb is an antonym of the positive matrix predicate in 7a. We labeled the condition 7b as "Matrix verb negative". For the majority of the stimuli items (12 items out of 16), the antonym verb in condition 7b happened to be formed by adding an overt negation marker to the positive predicate. The context for 7b is identical to 7a, but because the matrix verbs in these two conditions are antonyms, we expect the judgments provided to the target sentence should be the opposite. In this way we counterbalanced the the association between the True/False judgments and the high/low scope construals of the target sentences. In addition to the two ambiguous conditions in 7a and 7b, we also included unambiguous target sentences as control comparison conditions, see 7c and 7d. For the control conditions, only the high scope reading is grammatically available, because the lower embedding verb, e.g. *believe*, blocks the low scope dependency. The unambiguous target sentences were preceded by the same context used for the ambiguous conditions, resulting in the judgment *false* for 7c and *true* for 7d.

(7) Context: At a recent archaeology conference, Emily said that her research team found evidence to prove that a famous ancient city was actually built by aliens. But she kept the name of the city a secret. (Mandarin: 在最近的一次考古界的学术

会议上, 艾米丽说她的团队找到了证据证实某一个有名的古城市其实是外星人建造的。但目前她对这个城市的名字保密.)

Target sentence:

a. Ambiguous; Matrix verb positive

艾米丽 公布了 [$_{Clause1}$ 她的团队 发现了 [$_{Clause2}$外星人
Emily gongbu-le tade tuandui faxian-le waixingren
Emily **announce-perf** her team **discover-perf** aliens
建造了 哪座城市。]]
jianzao-le na-zuo chengshi
establish-perf which-CL-city

High scope: "Emily announced which city her team discovered aliens established." (False)

Low scope: "Emily announced her team discovered which city the aliens established." (True)

b. Ambiguous; Matrix verb negative

艾米丽 隐瞒了 [$_{Clause1}$她的团队 发现了 [$_{Clause2}$外星人
Emily yinman-le tade tuandui faxian-le waixingren
Emily **conceal-perf** her team **discover-perf** aliens
建造了 哪座城市。]]
jianzao-le na-zuo chengshi
establish-perf which-CL-city

High scope: "Emily concealed which city her team discovered aliens established." (True)

Low scope: "Emily concealed her team discovered which city the aliens established." (False)

c. Unambiguous; Matrix verb positive

艾米丽 公布了 [$_{Clause1}$她的团队 相信 [$_{Clause2}$外星人
Emily gongbu-le tade tuandui xiangxin waixingren
Emily **announce-perf** her team **believe** aliens
建造了 哪座城市。]]
jianzao-le na-zuo chengshi
establish-perf which-CL-city

High scope: "Emily announced which city her team believed aliens established." (False)

Low scope: “Emily announced her team believed which city the aliens established.” (Unavailable)

d. Unambiguous; Matrix verb negative

艾米丽 隐瞒了 [$_{Clause1}$她的团队 相信 [$_{Clause2}$外星人
Emily yinman-le tade tuandui xiangxin waixingren
Emily **conceal-perf** her team **believe** aliens
建造了 哪座城市。]]
jianzao-le na-zuo chengshi
establish-perf which-CL-city

High scope: “Emily concealed which city her team believed aliens established.” (True)

Low scope: “Emily concealed her team believed which city the aliens established.” (Unavailable)

We constructed a total of 16 sets of 4-condition items like 7a-d. The experiment was conducted on Ibex Farm (Drummond 2016). For each trial, participants first viewed a context scenario, and then they pressed the space bar to view the target sentence on the next screen. On the target sentence screen, they could not go back to view the context scenario. During the practice trials, between the context scenario and the target sentence there was a instruction sentence saying “You will next read a sentence. Please decide whether the meaning of that sentence matches or does not match the context scenario you just saw above”.[2]. They were instructed to decide, by choosing between two buttons presented to them on the screen. The 16 sets of experimental items were distributed to the participants in a Latin Square design, such that each participant only saw one condition from each item set. We also included 10 additional filler trials. The filler trials had the same format as the experimental trials, and 5 of them should be judged as true, while the other 5 as false. Ninety-eight native Mandarin speakers participated in our study, 10 of whom were excluded because their response accuracy on the filler trials was lower than 60%. We report the results from the remaining 88 participants below.

Results. We first converted participants’ truth value judgments into whether they interpreted the target sentence with a high scope construal. For example, for 7a, a response of *false* was converted to *high scope*; and for 7b, a response of *true* was converted to *high scope*. The proportion of high scope choices is plotted in Figure 1. There are more high scope responses for the unambiguous conditions (79% for the positive predicate condition and 77% for the negative predicate condition) than the ambiguous conditions (mixed effects logistic model: $Est = -0.21 \pm 0.08, z = -2.58, p < .01$). This is

unsurprising given that the unambiguous conditions can only be parsed as having a high scope for the wh-expressions. It is worth noting, however, that the proportions of high scope choices for unambiguously high-scope sentences like 7c and 7d are not at ceiling. As we will show in Experiment 2, the unambiguous conditions tested here are syntactically complex and received very low acceptability ratings. The severe parsing difficulty on the unambiguous conditions may have led to inaccurate interpretation sometimes. But we note that the primary interest of the current study is to explain the interpretation bias for the ambiguous conditions, and the interpretation of the unambiguous conditions does not play a major role for the main purpose of the paper. For the current purpose, the more important finding from Experiment 1 is that that the two ambiguous conditions both received overwhelmingly more high-scope responses, 73% for both the positive and the negative predicate conditions, significantly higher than the 50% chance level ($ps < .0001$).

< INSERT FIGURE 1 ABOUT HERE >

Discussion of Experiment 1. Results from the truth value judgment task in Experiment 1 provide strong evidence that participants are predominantly biased towards interpreting an ambiguous wh-in-situ construction as having a high scope reading. This finding is incompatible with the simple hypothesis that interpretation bias always tracks the parsing bias. As discussed in Section 1, there are good reasons to believe that from a parsing perspective, the local scope dependency (i.e. low scope) is less complex to establish and is the preferred parse for the parser, and the non-local scope dependency (i.e. high scope) is more complex and less preferred. The interpretation bias revealed by Experiment 1, however, is the opposite of the parsing bias.

This conclusion, that the interpretation bias obtained in Experiment 1 is the opposite of the parsing bias, critically depends on the assumption that there is a locality bias in parsing, which is based on previous findings in Xiang and Wang (2020). One potential concern is that although the constructions tested by Xiang and colleagues were the same as in the current study, the stimuli in the two studies are not exactly identical. We therefore conducted an acceptability judgment experiment in Experiment 2 to find out if the parsing locality bias would be replicated using the current set of stimuli.

2.3. Experiment 2: Reproducing the locality parsing bias in an acceptability rating task.

Material, participants, procedure and predictions. The Experiment material for Experiment 2 was identical to Experiment 1, with a total of 16 sets of 4-condition

experimental items (see an example in 7) and 10 filler items. The experimental procedure was also identical to Experiment 1 : each trial consisted of a context scenario followed by a target sentence. The only difference was that, instead of a truth value judgment task, at the target sentence participants were instructed to make a binary judgment (Yes/No) as to whether the target sentence was acceptable or not. Thirty native Mandarin speakers participated in the study. We excluded 6 participants whose accuracy on filler trials was below 60%. The data analysis reported below was based on the remaining 24 participants.

If there is a parsing bias favoring the local scope dependency, we make the following prediction. For the ambiguous conditions 7a and 7b, the local dependency is available, but for the unambiguous conditions 7c and 7d, the local dependency is blocked. The locality bias for the lower scope should manifest in a higher acceptability for the ambiguous than the unambiguous conditions, since in the latter case the favored low-scope parse is blocked and participants are forced to construct the disfavored high-scope parse. It is well known that parsing difficulty significantly reduces acceptability ratings (e.g. Chomsky & Miller 1963, Hofmeister et al. 2013).

Results and discussion. The acceptability judgment results support our prediction that there is a local scope preference. As shown in Fig. 2, the unambiguous conditions were rated significantly less acceptable (mean 0.44) than the ambiguous conditions (mean 0.67) regardless of whether the predicate was positive or negative ($Est = -1.01 \pm 0.37, z = -2.78, p < .01$)[3]. Sentences with positive matrix predicates were also rated lower than those with negative matrix predicates ($Est = -0.82 \pm 0.34, z = -2.44, p < 05$). Our focus here is not on the cause of the difference between the negative and positive so we won't discuss this at any length here. However, we note that since the context scenario in general ends on a sentence describing what did not happen, e.g. Emily kept the name of the city a secret in 7, this may have primed participants to favor a negative predicate over a positive one in the target sentence.

< INSERT FIGURE 2 ABOUT HERE >

The acceptability ratings for the ambiguous conditions 7a/b reflect a moderate degree of complexity for these sentences, whereas the much lower ratings for the unambiguous conditions 7c/d suggest severe parsing complexity. Both patterns are in line with the previous acceptability rating results on similar constructions[4]. The gradient acceptability in itself is not unusual for structurally complex sentences, given the well-established observations that even for completely grammatical sentences, heightened processing complexity can substantially reduce acceptability ratings (Chomsky & Miller 1963,

Hofmeister et al. 2013). Most important for the current purpose, the low ratings on the unambiguous conditions suggest that associating a wh-in-situ phrase with a non-local scope position is difficult and costly for comprehenders.

2.4. Summary of Experiment 1 and 2. The results from Experiment 1 and 2 present an empirical paradox. On the one hand, Experiment 2, using an acceptability rating task, confirmed the locality bias in parsing a wh-in-situ dependency. In particular, sentences that make a local scope dependency available are judged much more acceptable than sentences that block the local scope dependency and allow only a non-local high scope dependency. Experiment 1 with a truth value judgment task, however, showed an anti-locality bias in the ultimate interpretation participants obtained for sentences that are scope-ambiguous. The apparent contrast between the results from these two experiments sets up the core empirical observation that parsing biases do not necessarily align with interpretive biases. The parsing-interpretation misalignment revealed in Experiment 1 and 2 is broadly in line with previous findings based on "good-enough" misinterpretations, such as those found in garden-path or local-coherence sentences. In these other cases, it is possible for the comprehenders to obtain interpretations that are not licensed by the grammatical parse. In the current case, there is no "misinterpretation" per se, since the interpretations under consideration are all licensed by the grammar, but there is still a parsing-interpretation misalignment in the sense that at the global level the preferred interpretation is not the one compatible with the preferred parse.

The claim about the misalignment between parsing and interpretation, i.e. a locality bias for parsing and an anti-locality bias for interpretation, is only relevant for the ambiguous conditions tested in Experiment 1 and 2. The unambiguous conditions mainly serve as controls to help us detect the locality bias in parsing. The rest of the paper will therefore only focus on the ambiguous conditions, and we will develop a formal proposal to reconcile the discrepancies observed in Experiment 1 and 2. The main intuition we will develop is that sentence comprehension/interpretation should be modeled as the result of a pragmatic reasoning process between cooperative conversational partners. This is not to deny the role of parsing in sentence comprehension. In fact, comprehension requires constructing structural representations for the linguistic input, because semantic composition needs to consult the parsing outcome. In the current case study, a complete parse will specify where the scope position is for the wh-in-situ phrase. But in the meantime, parsing-based semantic composition is only the beginning but not the end of the interpretive process. In the rest of the paper, we explore the possibility that the currently observed contrast between parsing and interpretative decisions can be (at least

partly) captured by examining how a listener pragmatically reasons about the most likely messages the speaker has intended, given the possible parses of the utterance, and the listener's world knowledge.

The general idea that language communication should be viewed as a cooperative process between speakers and listeners, involving sophisticated pragmatic reasoning, is an old and extremely influential one (Grice 1975). In recent years, this insight has been formalized using bayesian pragmatic models, in particular the Rational Speech Act framework (RSA, Goodman & Frank 2016, Frank & Goodman 2012). Our proposal is built upon the RSA framework. In section 3, we first introduce some general background about the RSA model, and then extend the original model to the current case study. Most importantly, we will show that parsing preferences could be integrated with Bayesian pragmatic reasoning in a single model, and this makes the correct predictions, as shown by further empirical evaluations in section 4. Overall, our extended model successfully reconciles the apparent discrepancies between parsing and interpretation.

3. Integrating parsing biases with Bayesian pragmatic inferences.

3.1. The Rational Speech Act model of pragmatic inferences. The Rational Speech Act model (Goodman & Frank 2016, Frank & Goodman 2012) views speakers and listener as rational agents that collaborate on a language communication task. In a linguistic exchange, a listener and a speaker probabilistically and recursively reason about each other's behavior. A listener assumes that the utterance made by the speaker is meant to convey a particular state of the world (i.e. a message), with the understanding that the speaker chooses a particular utterance instead of any other alternatives because they reason about how an utterance would be interpreted by a listener. The recursive reasoning could continue on for many levels of iterations, but minimally we could consider three levels for the current purpose: a pragmatic listener, a pragmatic speaker and a literal listener. On the top is a level of inference of a pragmatic listener. Upon hearing an utterance, a pragmatic listener would update his probabilistic model of the world states based on the information conveyed by the utterance. A pragmatic listener's posterior belief about a particular world state w given the utterance u, using the Bayes rule, is shown in equation 8:

$$P_L(w|u) = \frac{P_S(u|w) \times P(w)}{\sum_{w'} P_S(u|w') \times P(w')} \tag{8}$$

The pragmatic listener (L) conditions his belief update on two factors. First, assuming the speaker S is cooperative and trying to be helpful, the listener works backwards and estimates the likelihood a speaker would have uttered u given the world state w in the

speaker's mind (the term $P_S(u|w)$). Second, the listener also brings to the communication his prior belief as to how likely the world state w holds independent of the utterance (the term $P(w)$[5]). The normalizing constant in 8 (i.e. the denominator) considers the alternative world states that could have been relevant in the communicative context.

The inferences of a pragmatic speaker, i.e. the term $P_S(u|w)$ in equation 8, is defined in the following way. A speaker could have more than one choice of utterances when she linguistically encodes a particular world state, but her decision was assumed to be rational: she chooses her utterance from a set of alternative utterances according to the utility U_s that a particular utterance would obtain, as shown in 9. A rational pragmatic speaker would in general want to maximize her utility, and the free parameter α in 9 captures the extent to which the speaker is a rational agent, i.e. how much she would choose her utterance to maximize her utilities. When $\alpha = 0$, the speaker's choices are random; but as $\alpha \to \infty$, the speaker chooses the utterance with the greatest utility. The utility function could be defined in a number of ways (Goodman & Frank 2016), and we follow the most basic definition that states a pragmatic speaker would choose to make the most informative utterance to the listener, as shown in 10[6]. To avoid infinite recursion, the listener in 10 is defined to be a simple *literal* listener L_0. Based on the equation in 10, utterances with high utility are those that would make the literal listener boost the probability of the world state w intended by the speaker. The literal listener updates his probabilistic beliefs about different world states (i.e. the term $P_{L_0}(w|u)$) based on whether the literal meaning (i.e. the semantic meaning) of the utterance is compatible with the relevant world states or not, as shown in 11.

$$(9) \quad P_S(u|w) \propto exp(\alpha \times U_S(u;w))$$

$$(10) \quad U_S(u;w) = ln(P_{L_0}(w|u))$$

$$(11) \quad P_{L_0}(w|u) = \frac{\delta_{[\![u]\!](w)} P(w)}{\sum_{w' \in W} \delta_{[\![u]\!](w')} P(w')}$$

The literal listener's inference in 11 is crucial for the standard RSA model – this is the level at which the compositional semantics of the linguistic input is imported into the pragmatic reasoning process. The term $\delta_{[\![u]\!](w)}$ in 11 takes the value 1 or 0, determined by whether the utterance $[\![u]\!]$ is compatible or not with a given world state w. All the world states that will make the utterance false will be removed, and the literal listener will update their beliefs based on the remaining world states (i.e. the ones that are compatible with the semantics of the utterance). Since this is the place compositional semantics meets pragmatic reasoning, we propose that parsing biases could be incorporated into the pragmatic reasoning process at the listeral listener's level. In particular, the basic

form of the literal listener in 11 deals with simple unambiguous utterances. With a more complex ambiguous utterance u, if it has n possible structural parses that partition the entire parsing space (i.e., the probabilities of these parses add up to 1), it can be proved that the literal listener's inference about u is the sum of L_0's inference of each possible parse weighted by the probability of that parse. We demonstrate this extended version of L_0's inference in 12. Pertaining to the empirical case of our interest, we assume the utterance u in 12 has two possible parses, u_h and u_l, representing a wh-high-scope and a wh-low-scope parse respectively[7]. The probability of a world state is first computed for each structural parse separately by applying 11 to that parse, and then the information was summed together after being weighted by the probability of each parse.

(12) $P_{L_0}(w|u)$

$$= P_{L_0}(w|u_h) \times P(u_h) + P_{L_0}(w|u_l) \times P(u_l)$$

$$= \frac{\delta_{[\![u_h]\!](w)} P(w)}{\sum_{w'} \delta_{[\![u_h]\!](w')} P(w')} \times P(u_h) + \frac{\delta_{[\![u_l]\!](w)} P(w)}{\sum_{w'} \delta_{[\![u_l]\!](w')} P(w')} \times P(u_l)$$

In the rest of this section, we will apply equations 8 - 12 to understand the empirical puzzle raised in Experiment 1 and 2. A wh-in-situ utterance with scope ambiguity could in principle be used to convey a number of different states of the world. The ultimate task for us is to derive, based on 8, a pragmatic listener's posterior probability for each relevant world state upon hearing an ambiguous utterance like the one in 7a/b. We will do this in section 3.6 in a qualitative manner, and then later in section 4.2 with more quantitative measures. Also in section 4.2, we will link the posterior probabilities of a pragmatic listener to the empirical truth value judgments result obtained in Experiment 1. But prior to applying 8, we first need to implement a number of other necessary steps and work through equations 9-12. First, in section 3.2, we will define the relevant world states for an ambiguous wh-in-situ utterance. Next, we will experimentally estimate the prior term $P(w)$ for each world state in section 3.3. Then in section 3.4 we derive the literal listener's inference based on equation 11 and 12. This is a crucial step since the parsing bias of the wh-in-situ utterances will be integrated with the RSA model at this step. And after that in section 3.5 we derive the speaker inference based on equation 9 and 10. Finally in section 3.6 we put everything together and derive the pragmatic listener's inference based on equation 8.

3.2. Defining the relevant world states. If a listener's interpretation process is modeled as updating her beliefs of the relevant world states w given an utterance u, it is important to be clear what the relevant world states could be for the current case study.

Our main interest is the ambiguous wh-in-situ utterances u in 7a and 7b. These examples are repeated below in 13a and 13b:

(13) a. Ambiguous; Matrix verb positive

艾米丽 公布了 [$_{Clause1}$ 她的团队 发现了 [$_{Clause2}$ 外星人
Emily **announce-perf** her team **discover-perf** aliens
建造了 哪座城市。]]
establish-perf which-CL-city

High scope: "Emily announced which city her team discovered aliens established."

Low scope: "Emily announced her team discovered which city the aliens established."

b. Ambiguous; Matrix verb negative

艾米丽 隐瞒了 [$_{Clause1}$ 她的团队 发现了 [$_{Clause2}$ 外星人
Emily **conceal-perf** her team **discover-perf** aliens
建造了 哪座城市。]]
establish-perf which-CL-city

High scope: "Emily concealed which city her team discovered aliens established."

Low scope: "Emily concealed her team discovered which city the aliens established."

These utterances are ambiguous, and could convey information about different world states. The high or low-scope readings of the sentences above are semantic meanings derived from particular structural representations (i.e. depending on the scope dependency), and in principle, each of them could be compatible with one or more states in the world. Let's first make clear what the most relevant world states could be for our working example in 13. When the matrix predicate is positive, as in 13a, the relevant world states are a set of possible combinations of two events: e1: Emily announced the name of a city, which her team discovered was built by aliens; and e2: Emily announced their discovery that there was a city that was built by aliens. Let's call e1 the *name announcement* event, and e2 the *discovery announcement* event. There are a total of 4 different ways to combine these two events, assuming each event takes either a *true (+)* or *false (−)* value, as shown in Table 1. Out of the 4 combinations, w_2 is not logically possible, since Emily couldn't have announced the name of the city that they discovered was built by aliens without also announcing that they made such a discovery. In addition,

w_4 is irrelevant since if neither event is true, the speaker wouldn't have uttered 13a in the first place. The two remaining world states w_1 and w_3 are therefore the two relevant states the pragmatic listener considers for the target sentence she hears. Applying the same reasoning to the target sentence with a negative matrix predicate, as in 13b, the relevant world states are also a set of possible combinations of two events: the *name concealing* event e1: Emily concealed the name of a city, which her team discovered was built by aliens; and the *discovery concealing* event e2: Emily concealed their discovery that there was a city that was built by aliens. Among the 4 combinations of these two events, shown in Table 2, w_3 is logically impossible, because one can not conceal the discovery of the city without also concealing the name of the city that was discovered. The possibility w_4 in Table 2 is again trivially irrelevant. The pragmatic listener would therefore consider two relevant world states w_1 and w_2 in Table 2 upon hearing the target sentence.

< INSERT TABLE 1 ABOUT HERE >

< INSERT TABLE 2 ABOUT HERE >

In Table 3, we summarize the remaining relevant world states considered by the listener given the target sentences. The remaining relevant world states are relabeled in Table 3 as w_1 and w_2, and these are the w_1 and w_2 we will refer to in the later discussion. Note that for the positive and negative utterances, their corresponding w_2 states are essentially representing identical world affairs; but their corresponding w_1 states are different.

< INSERT TABLE 3 ABOUT HERE >

With the relevant world states defined as above, in the next section we experimentally estimate the prior probability for each state, i.e. $P(w_1)$ and $P(w_2)$.

3.3. Experiment 3: Estimating the prior probabilities.

Material, participants and procedure. To experimentally assess the prior probabilities of each different world state relevant to the listener, we first provided participants a neutral context that corresponds to the background scene used in the truth value judgment task in Experiment 1, for example, a background scene about an archaeology conference[8]. Once participants viewed the context sentence, on the next screen participants were instructed to choose between two possible situations that could

take place in the given context. These two situations correspond to the two different world states illustrated in Table 3 (with different paraphrases). World states for sentences with positive and negative matrix predicates were tested in two different conditions in a within-subject design. The experiment material was closely modeled after material from Experiment 1. Sixteen sets of items corresponding to the original 16 sets of scenarios in Experiment 1 were constructed, with two conditions in each set of items. Each condition contains two choices (w_1 and w_2). An example item set is given in 14 below. As we mentioned earlier in Table 3, the w_2 states under the positive and negative matrix predicates represent identical world affairs. We therefore used identical paraphrases for the w_2 situation under 14a and 14b. In Figure 3 we also present an example trial.

(14) Context: At a recent archaeology conference, Emily made a presentation on behalf of her research team. (Mandarin: 在最近的一次考古界的学术会议上, 艾米丽代表她的研究团队作了一个报告。)
Question: Which of the following situation is more likely to arise? (Mandarin: 以下的哪种情况更有可能发生？)

a. The positive predicate condition:
w_1: In her report, Emily said that her research team found evidence to prove that a famous ancient city was actually built by aliens. She also released the name of the city. (Mandarin: 在她的报告里，艾米丽说她的团队找到了证据证实某一个有名的古城市其实是外星人建造的，她同时也宣布了这个城市的名字。)
w_2: In her report, Emily said that her research team found evidence to prove that a famous ancient city was actually built by aliens. But the name of the city needs to be kept secret for the moment. (Mandarin: 在她的报告里，艾米丽说她的团队找到了证据证实某一个有名的古城市其实是外星人建造的，但目前她需要对这个城市的名字保密。)

b. The negative predicate condition:
w_1: Emily's research team actually has found evidence to prove that a famous ancient city was built by aliens. But in her report she didn't mention this discovery at all. (Mandarin: 艾米丽的团队其实已经找到了证据证实某一个有名的古城市是外星人建造的，但她在自己的报告里完全隐瞒了这个发现.)
w_2: In her report, Emily said that her research team found evidence to prove that a famous ancient city was actually built by aliens. But the name of the city needs to be kept secret for the moment. (Mandarin: 在她的报告里，艾米丽说她的团队找到了证据证实某一个有名的古城市其实是外星人建造的，但目

前她需要对这个城市的名字保密。）

< INSERT FIGURE 3 ABOUT HERE >

The experiment was conducted on IbexFarm. A hundred and nineteen native Mandarin speakers participated in our study. The 16 sets of experimental items were distributed to participants with a Latin Square distribution, such that each participant only saw one of the two conditions for each item. There were also an additional 10 filler items, so each participant finished a total of 26 trials.

Results. Among the choices participants made for the *positive predicates* condition, there was on average a slight numerical preference for the w_1 state (0.53 w_1 vs. 0.47 w_2), but it was not different from chance ($p = 0.07$); for the *negative predicates* condition, there was a preference for w_2 over w_1 (0.42 w_1 vs. 0.58 w_2), significantly different from chance ($p < .0001$).

3.4. Connecting parsing outcomes to pragmatic reasoning. With the world states defined and the prior probability of each state estimated, in this section we demonstrate how to integrate parsing biases and pragmatic reasoning into a single model. In particular, we will work through equations 11 and 12 in this section. As mentioned in section 3.1, a full bayesian pragmatic model carries out recursive reasoning between a listener and a speaker. A pragmatic speaker makes decisions about their production choices by reasoning about the linguistic update of a literal listener, and the outcome from the pragmatic speaker stage is in turn used to update a pragmatic listener's inferences. As the starting point of this chain of reasoning, the literal listener L_0 is the crucial step that connects structured semantic composition to pragmatic reasoning. The literal listener L_0 does this by performing a belief update about different world states based on the literal meaning of a heard utterance. The basic formulation of L_0 in equation 11, adapted from the original RSA framework, only applies to utterances that are structually simple and unambiguous. Extending it to deal with syntactically complex and ambiguous sentences, we make the simple assumption that the compositional semantics of an utterance u depends on how the surface string is parsed. In the current case, a target wh-in-situ sentence has two possible parses, each representing one type of scope dependency. Let's call the two parses u_h and u_l, standing for the high-scope parse and the low-scope parse. As shown in 12, repeated in 15, we can calculate the L_0's inferences for an ambiguous utterance by combining different parses based on the probability of each parse:

(15) $P_{L_0}(w|u)$

$$= P_{L_0}(w|u_h) \times P(u_h) + P_{L_0}(w|u_l) \times P(u_l)$$

$$= \frac{\delta_{[\![u_h]\!](w)} P(w)}{\sum_{w'} \delta_{[\![u_h]\!](w')} P(w')} \times P(u_h) + \frac{\delta_{[\![u_l]\!](w)} P(w)}{\sum_{w'} \delta_{[\![u_l]\!](w')} P(w')} \times P(u_l)$$

To see how 15 applies to the current empirical case, let's consider our working example in 13a, in which the matrix predicate is a positive predicate. The English glosses for 13a are repeated in 16. For convenience, we also repeat from Table 3 the two world states relevant for this utterance.

(16) Emily *announced* her team *discovered* aliens established which city.
(艾米丽公布了她的团队发现了外星人建造了哪座城市。)
High scope parse: "Emily announced which city her team discovered aliens established."
Low scope parse: "Emily announced her team discovered which city the aliens established."
w_1 *positive*: Emily announced they discovered that a city was built by aliens and she also announced the name of the city.
w_2 *positive*: Emily announced they discovered that a city was built by aliens but she did not announce the name of the city.

Based on 15, we can compute the posterior probabilities a literal listener has for the world state w_1 and w_2 upon hearing the ambiguous utterance in 16. To start with, we will first make the simple assumption that it is equally likely for a literal listener to parse the ambiguous string in 16 into a high-scope or a low-scope dependency, i.e. $P(u_h)$ and $P(u_l)$ are equal at 0.5. We know this is in fact not true, since there is a locality bias in parsing that favors the low-scope parse (see Experiment 2 and the discussion there), and we will come back to modify this assumption at the end of this section. If the utterance u in 16 is parsed as u_h, it specifies the fact that the name of the city was made known. Under this parse the utterance is compatible with w_1 (hence $\delta_{[\![u_h]\!](w_1)}$=1 in 17 but incompatible with w_2 (hence $\delta_{[\![u_h]\!](w_2)}$=0). If the utterance u is parsed as u_l, since it underspecifies whether the name of the city is made known, it is compatible with both w_1 and w_2. We therefore cannot remove either w_1 or w_2 from consideration, and both are kept as viable options for the listener to consider. In addition, we already know the prior probabilities for $P(w_1)$ and $P(w_2)$ are 0.53 and 0.47 (see Experiment 3). The literal listener L_0 therefore updates her beliefs about w_1 and w_2 in the following way:

(17) a. $P_{L_0}(w_1|u_{positive})$

$$= \frac{\delta_{[\![u_h]\!](w_1)} P(w_1)}{\delta_{[\![u_h]\!](w_1)} P(w_1) + \delta_{[\![u_h]\!](w_2)} P(w_2)} \times P(u_h) +$$

$$\frac{\delta_{[\![u_l]\!](w_1)}P(w_1)}{\delta_{[\![u_l]\!](w_1)}P(w_1)+\delta_{[\![u_l]\!](w_2)}P(w_2)} \times P(u_l)$$

$$= \frac{1 \times 0.53}{1 \times 0.53 + 0 \times 0.47} \times 0.5 + \frac{1 \times 0.53}{1 \times 0.53 + 1 \times 0.47} \times 0.5$$

$$= 1 \times 0.5 + 0.53 \times 0.5$$

$$= 0.765$$

b. $P_{L_0}(w_2|u_{positive})$

$$= \frac{\delta_{[\![u_h]\!](w_2)}P(w_2)}{\delta_{[\![u_h]\!](w_1)}P(w_1)+\delta_{[\![u_h]\!](w_2)}P(w_2)} \times P(u_h) +$$

$$\frac{\delta_{[\![u_l]\!](w_2)}P(w_2)}{\delta_{[\![u_l]\!](w_1)}P(w_1)+\delta_{[\![u_l]\!](w_2)}P(w_2)} \times P(u_l)$$

$$= \frac{0 \times 0.47}{1 \times 0.53 + 0 \times 0.47} \times 0.5 + \frac{1 \times 0.47}{1 \times 0.53 + 1 \times 0.47} \times 0.5$$

$$= 0 + 0.47 \times 0.5$$

$$= 0.235$$

The results from the calculation in 17 suggests that even though the literal listener starts with a prior belief that the probabilities for w_1 and w_2 are very close to each other (0.53 and 0.47), after hearing the utterance in 16, the literal listener is leaning much more towards believing in w_1 over w_2.

The working example from 13b, in which the utterance contains a negative matrix predicate, is repeated in 18. The calculation in 19 is very similar to the positive predicate case in 17, but the compatibility between the utterance and each world state changes. When the utterance u is parsed as u_h, it is compatible with both w_1 and w_2, hence both states need to be considered by the listener. If the utterance is parsed as u_l, it is only compatible with w_1, and w_2 will be removed from further consideration. In addition, the prior probabilities for w_1 and w_2 were estimated to be 0.42 and 0.58 from Experiment 3.

(18) Emily *concealed* her team *discovered* aliens established which city.
(艾米丽隐瞒了她的团队发现了外星人建造了哪座城市.)
High scope parse: "Emily concealed which city her team discovered aliens established."
Low scope parse: "Emily concealed her team discovered which city the aliens established."

w_1 *negative*: Emily concealed the fact that they discovered that a city was built by aliens and also concealed the name of the city.
w_2 *negative*: Emily did not conceal the fact that they discovered that a city was built by aliens, but she concealed the name of the city.

(19) a. $P_{L_0}(w_1|u_{negative})$

$$= \frac{\delta_{[\![u_h]\!](w_1)}P(w_1)}{\delta_{[\![u_h]\!](w_1)}P(w_1) + \delta_{[\![u_h]\!](w_2)}P(w_2)} \times P(u_h) +$$

$$\frac{\delta_{[\![u_l]\!](w_1)}P(w_1)}{\delta_{[\![u_l]\!](w_1)}P(w_1) + \delta_{[\![u_l]\!](w_2)}P(w_2)} \times P(u_l)$$

$$= \frac{1 \times 0.42}{1 \times 0.42 + 1 \times 0.58} \times 0.5 + \frac{1 \times 0.42}{1 \times 0.42 + 0 \times 0.58} \times 0.5$$

$$= 0.42 \times 0.5 + 1 \times 0.5$$

$$= 0.71$$

b. $P_{L_0}(w_2|u_{negative})$

$$= \frac{\delta_{[\![u_h]\!](w_2)}P(w_2)}{\delta_{[\![u_h]\!](w_1)}P(w_1) + \delta_{[\![u_h]\!](w_2)}P(w_2)} \times P(u_h) +$$

$$\frac{\delta_{[\![u_l]\!](w_2)}P(w_2)}{\delta_{[\![u_l]\!](w_1)}P(w_1) + \delta_{[\![u_l]\!](w_2)}P(w_2)} \times P(u_l)$$

$$= \frac{1 \times 0.58}{1 \times 0.42 + 1 \times 0.58} \times 0.5 + \frac{0 \times 0.58}{1 \times 0.42 + 0 \times 0.58} \times 0.5$$

$$= 0.58 \times 0.5 + 0$$

$$= 0.29$$

The observation here is that even though the literal listener started with a lower prior probability for w_1 (0.42), after hearing the utterance, the listener's posterior beliefs have changed to favor w_1 over w_2 (0.71 vs. 0.29).

The calculations in 17 and 19 showed that a literal listener, upon hearing a scope-ambiguous wh-sentence, would favor w_1 over w_2 regardless of whether the predicate is positive or negative. The calculations so far are based on the assumption that the literal listener has no parsing bias while parsing an ambiguous string u into either a high-scope or a low-scope dependency ($p(u_h) = p(u_l) = 0.5$). This assumption needs refinement, since we already know the parser favors the low scope dependency over the high scope one. After we introduce the constraint $0 < p(u_h) < 0.5$ and $0.5 < p(u_l) < 1$ into

the calculations in 17 and 19, we will derive that for utterances with a positive predicate like 17, the literal listener's posterior probability for w_1 is between 0.53 and 0.765; and for utterances with a negative predicate like 19, it is between 0.71 and 1. In other words, upon hearing a scope-ambiguous target utterance, given the parsing preference that favors the low-scope dependency, the literal listener is predicted to assign higher posterior probability to w_1 than w_2, regardless of the polarity of the predicate.

3.5. From the literal listener to the pragmatic speaker. With the inferences of a literal listener, we can now model the next level of inferences: the pragmatic speaker's inferences in equations 9 and 10. These two equations are combined and presented/repeated in 20. According to 20 the speaker's choice of an utterance is largely determined by the informativity of this utterance. The probability of a speaker choosing an utterance to convey a world state is proportional to the posterior probability that the literal listener L_0 infers about the target world state upon hearing that utterance.

(20) $P_S(u|w) \propto exp(\alpha \times ln(P_{L_0}(w|u))$

The pragmatic speaker makes their production choices by comparing the informativity/utility of all the alternative utterances for a given world state. The contribution of the alternative utterances can be seen more clearly in 21, which includes the proportionality constant that is not specified in 20.

(21) $P_S(u|w)$

$$= \frac{exp(\alpha \times U_S(u;w))}{\sum_{u' \in ALT} exp(\alpha \times U_S(u';w))}$$

$$= \frac{exp(\alpha \times ln(P_{L_0}(w|u))}{\sum_{u' \in ALT} exp(\alpha \times ln(P_{L_0}(w|u'))}$$

Based on 21, given a well-defined set of alternative utterances $u_1, u_2, ...u_n$ and a set of relevant world states $w_1, w_2, ...w_k$, one can calculate the production likelihood $P(u_i|w_j)$ for each pair of u and w. One critical question is how to define the set of alternative utterances available to a speaker. Previous studies using the basic RSA framework often investigate syntactically simple structures, and it is relatively straightforward to define the set of alternative utterances for a speaker. For example, in the case of quantity implicature calculation that derives the *some but not all* inference from the quantifier *some*, it is reasonable to hypothesize that, *some* and *all* form the set of alternative expressions that a speaker could choose from. In order to apply 21 to the current empirical case, however,

there are a number of practical challenges. In particular, the target structure of interest in the current study, the multi-clausal wh-in-situ construction, is much more complex[9]. It is difficult to define in advance, on a principled ground, the possible alternative structures a speaker may use. We conducted a production experiment to have a better assessment. The details of this experiment (Experiment 4) will be introduced in the next section. Overall, the empirical production results revealed to us a nuanced set of structures from participants, which can also vary depending on the context scenarios and the lexical items involved. Informed by the empirical production results, we make the following simplifying assumptions. Since a large number of the alternative structures produced by participants in our production experiment are unambiguous, we take the unambiguous utterances as the major type of alternative choice a speaker has. We therefore assume three types of utterances available to the pragmatic speaker: the ambiguous target wh-in-situ construction u_{ambig}, which is compatible with different world states; the utterance $u_{unambig1}$ that unambiguously describes the world state w_1 and is therefore incompatible with w_2; and finally the utterance $u_{unambig2}$ that unambiguously describes the world state w_2 and is incompatible with w_1. A literal listener's update based on these three types of utterances is presented in Table 4.

< INSERT TABLE 4 ABOUT HERE >

Given the three types of alternative utterances in Table 4, which are an approximation but not a precise representation of all the possible utterances, we derive the probability of a pragmatic speaker choosing the target wh-in-situ form u_{ambig} to describe w_1 and w_2 in the following way (based on 21:

(22) $P_S(u_{ambig}|w_1)$

$$= \frac{exp(\alpha \times ln(P_{L_0}(w_1|u_{ambig}))}{\sum_{u' \in ALT} exp(\alpha \times ln(P_{L_0}(w_1|u'))}$$

$$= \frac{exp(\alpha \times ln(P_{L_0}(w_1|u_{ambig}))}{exp(\alpha \times ln(P_{L_0}(w_1|u_{ambig})) + exp(\alpha \times ln(P_{L_0}(w_1|u_{unambig1})) + exp(\alpha \times ln(P_{L_0}(w_1|u_{unambig2}))}$$

$$= \frac{(P_{L_0}(w_1|u_{ambig}))^\alpha}{(P_{L_0}(w_1|u_{ambig}))^\alpha + 1 + 0}$$

$$= \frac{(P_{L_0}(w_1|u_{ambig}))^\alpha}{(P_{L_0}(w_1|u_{ambig}))^\alpha + 1}$$

(23) $P_S(u_{ambig}|w_2)$

$$= \frac{exp(\alpha \times ln(P_{L_0}(w_2|u_{ambig}))}{\sum_{u' \in ALT} exp(\alpha \times ln(P_{L_0}(w_2|u'))}$$

$$= \frac{exp(\alpha \times ln(P_{L_0}(w_2|u_{ambig}))}{exp(\alpha \times ln(P_{L_0}(w_2|u_{ambig})) + exp(\alpha \times ln(P_{L_0}(w_2|u_{unambig1})) + exp(\alpha \times ln(P_{L_0}(w_2|u_{unambig2}))}$$

$$= \frac{(P_{L_0}(w_2|u_{ambig}))^\alpha}{(P_{L_0}(w_2|u_{ambig}))^\alpha + 0 + 1}$$

$$= \frac{(P_{L_0}(w_2|u_{ambig}))^\alpha}{(P_{L_0}(w_2|u_{ambig}))^\alpha + 1}$$

Because we already know from the last section that $P_{L_0}(w_1|u_{ambig}) > P_{L_0}(w_2|u_{ambig})$ regardless of whether the target utterance contains a positive or a negative predicate, it can be derived that $P_S(u_{ambig}|w_1) > P_S(u_{ambig}|w_2)$[10]. That is to say, the pragmatic speaker is more likely to use a scope ambiguous wh-in-situ target utterance when describing w_1 than when describing w_2 (regardless of whether the target utterance contains a positive or a negative predicate). We will come back to this prediction in section 4.1, in which we present a production experiment to test this prediction.

3.6. Putting everything together – deriving the pragmatic listener' inferences. We are now ready to tackle the inferences made by a pragmatic listener. Using the Bayesian inference rule in 8, the following relations in 25 holds: a pragmatic listener's posterior probability of a world state conditioned on a target wh-in-situ utterance is proportional to the product of the probability of a speaker choosing that utterance to convey the target world state and the prior probability of that world state.

(24) a. $P_L(w_1|u) \propto P_S(u|w_1) \times P(w_1)$

b. $P_L(w_2|u) \propto P_S(u|w_2) \times P(w_2)$

(25) a. $P_L(Target|Adj) \propto P_S(Adj|Target) \times P(Target)$

b. $P_L(Competitor|Adj) \propto P_S(Adj|Competitor) \times P(Competitor)$

We have empirically estimated the prior probabilities of the relevant world states w_1 and w_2, and we have also derived the speaker probabilities in the last section, which in turn were based on the literal listener's inferences that take into account the parsing biases.

Let's first consider the situation in which the utterance u is an ambiguous wh-in-situ construction with a positive predicate (see an example in 13a. For the prior probabilities,

based on Experiment 3, for utterances with a positive predicate, $P(w_1) > P(w_2)$ (0.53 w_1 vs. 0.47 w_2). For the speaker probabilities, we know from the last section that $P_S(u|w_1) > P_S(u|w_2)$. Combining these information with 25, we can conclude the following in 26: upon hearing the utterance in 13a, the pragmatic listener has a higher posterior probability for the world state w_1 than the world state w_2.

(26) $P_L(w_1|u) > P_L(w_2|u)$

But the situation is less straightforward when the utterance u contains a negative predicate (see an example in 13b. On the part of the speaker probabilities, based on the discussion in the last section we still have $P_S(u|w_1) > P_S(u|w_2)$; but the prior probabilities estimated from Experiment 3 revealed $P(w_1) < P(w_2)$ (0.42 w_1 vs. 0.58 w_2). Given these, the specific relation between $P_L(w_1|u)$ and $P_L(w_2|u)$ is uncertain: according to 25, whether the pragmatic listener assigns a higher posterior probability to w_1 or to w_2 (i.e. whether $P_L(w_1|u) > P_L(w_2|u)$) depends on the magnitude of the difference between the two speaker probabilities $P_S(u|w_1)$ and $P_S(u|w_2)$. One way to make a more precise assessment of these two speaker probabilities is to define a specific value for the free parameter α in equations 22 and 23. To avoid making arbitrary decisions on free parameter values, in the next section we conduct a production experiment to obtain an empirical estimate of the speaker probabilities. As we will also show below, an additional advantage for collecting empirical production data is that it also helps us test an independent prediction made at the end of section 3.5.

To summarize, in section 3 we developed an analysis that incorporates the parsing biases into the pragmatic reasoning process. In particular, applying the RSA model to the current empirical case, we demonstrated that parsing biases could be integrated into the literal listener's inferences in a principled fashion, which were then ultimately integrated into the pragmatic listener's inferences via an intermediate level of speaker inferences. The step-by-step demonstration in this section provides a detailed outline of the general proposal. The proposal makes clear qualitative predictions about the pragmatic speaker's posterior probabilities for utterances containing positive predicates. The predictions for utterances containing negative predicates are left open since a precise prediction would depend on a more specific estimate of the speaker inferences. In the next section, we conduct a production experiment to empirically estimate the speaker inferences. The goal of this production experiment is two-fold. First, with the empirically estimated speaker probabilities, we can empirically derive the pragmatic listener's posterior probabilities using the Bayes rule in 8, and we will then be able to evaluate whether the posterior probabilities of a pragmatic listener correctly predict the truth value judgment results

obtained from Experiment 1. Second, the empirical production results also allow us to validate a crucial prediction of the proposal: as a result of integrating parsing biases into the literal listener's inferences, in section 3.5 we derived a prediction that a pragmatic speaker is more likely to use a scope ambiguous wh-in-situ utterance for conveying w_1 than w_2. We will evaluate whether this prediction is borne out in the empirical data.

4. Empirically deriving the pragmatic listener's inferences.

4.1. Experiment 4: estimating the production pattern.

Material, procedure and participants. The goal of this experiment is to estimate how likely participants are to use the target wh-in-situ construction to describe a given world state. To this end, we first constructed scenarios that correspond to the four types of world states presented in Table 3. Next, we elicited productions that describe these world state scenarios. In particular, we are interested in whether participants will produce utterances identical or very similar to the ambiguous wh-in-situ target sentences used in the truth value judgment task in Experiment 1, as in 7a and 7b. One methodological concern is that the target wh-in-situ construction is complex, and it is very unlikely that a free production task will trigger sufficient (or any) amount of target production. Previous production results from Xiang and colleagues (2015) showed that native Mandarin speakers avoid producing relatively long wh-in-situ dependencies as much as they can, even at the cost of producing some otherwise dispreferred complex clause structures (e.g. relative clauses). Given this constraint, instead of eliciting free production, we provided phrase fragments to guide and constrain the participants' production process.

We constructed a total of 16 item sets, with each item set containing 4 conditions, corresponding to the 4 relevant world states. The experimental trials have the following structure. Participants saw one of the four world state scenarios on the screen. The phrasing of these scenarios were adapted and modified from the material used in Experiment 3 (see an example in 14. Below the scenario on the same screen, participants saw four phrase fragments. The participants were instructed to form a sentence using these fragments (by typing out a sentence that included these fragments), which expresses a message consistent with the scenario presented to them. The four fragments were presented in a 2x2 grid format, and the position of each fragment in the grid was randomized from trial to trial. For example, if a participant received a world state scenario for a positive predicate, e.g. either one of the two world states under 14a, the four fragments they would receive were *"Emily announced"*, *"which city"*, *"established"*, *"her team discovered"*. The same set of fragments were supplied to the participants for

both the w_1 and the w_2 scenarios under the same positive predicate. If a participant received a relevant world state scenario for a negative predicate, e.g. either one of the two world states under 14b, they would receive an almost identical set of fragments as above except that the positive predicate *"Emily announced"* is replaced by a negative one *"Emily concealed"*. The positions of these fragments in the 2x2 grid were randomized so that participants were not cued about the word order of the target sentence they were about to produce. An example trial is given in figure 4. During the practice trials participants were instructed that they could also add other material they want to use, as long as they include the provided phrases in their production. Even though the task itself is not equivalent to spontaneous natural production, it nevertheless leaves participants enough flexibility to form various types of utterances. Therefore they were not forced to produce the target structure. The experiment material was adapted from Experiment 1 and Experiment 3. The world state scenarios were adapted from Experiment 3 (e.g. example 14, and the phrase fragments were adapted from the target sentences in Experiment 1. The experiment was conducted on IbexFarm, and participants typed up and submitted each sentence they formed. A total of 248 native Mandarin speakers participated in our study. Each participant performed the task on 16 experimental trials and an additional 10 filler trials.

< INSERT FIGURE 4 ABOUT HERE >

Results. Three different native Mandarin speakers coded the production results. We removed the trials from participants that didn't perform the task properly (about 1% of the total trials). For each trial, if the participant produced a wh-in-situ structure similar to the target sentence in the truth value judgment task in Experiment 1, it was coded as a target structure. Similarity was evaluated based on whether the four fragments provided to the participants were organized into the same word order and syntactic structure as the target sentences in Experiment 1. All other structures they produced were coded as non-target structures.

On average, about 40% of the total trials conformed to the wh-in-situ target structure, with similar word order as the target sentences used in Experiment 1. In Figure 5, we present the proportion of the target structure produced, split by the world state context and the predicate type. Importantly, the results presented here confirm the prediction we made in section 3.5: for both types of predicates, participants were more likely to produce the ambiguous target structure when describing the w_1 state than the w_2 state (positive predicate: w_1 48%, w_2 31%; negative predicate: w_1 44%, w_2 36%), as confirmed

by a significant main effect of world state ($Est = 0.32 \pm 0.05, z = 6.8, p < .00001$)[11].

< INSERT FIGURE 5 ABOUT HERE >

It is worth noting that the target wh-in-situ dependency structure was not frequently produced by the participants (about 40% on average). This is not surprising given that the long wh-in-situ dependency is syntactically complex. Among the alternative structures participants produced, the most common strategy to reduce complexity was to produce unambiguous conjoined clauses such as "Emily announced her team discovered there was a city that was built by aliens, but she didn't announce which city it was (艾米丽公布了她的团队发现有一个古城市是外星人建造的，但是她没有公布是哪座城市.)". The conjoined-clause structure is longer than the target structure, but the scope of the wh-phrase is not ambiguous. On a small number of trials participants also produced structures that, although still scope-ambiguous, contained shorter scope-dependencies, such as "Emily announced her team discovered which city was built by aliens (爱丽丝公布了她的团队发现了哪座城市是外星人建造的.)". Compared to the target structure, this example employs shorter scope-dependencies due to the fact that the wh-phrase is in a clause-initial position instead of a clause-final position – both orders are grammatical in Mandarin. In addition, this structure had a different kind of information structure packaging compared to the target structure.[12].

4.2. Empirically deriving the pragmatic listener's inferences. We are ready to work out the the pragmatic listener's inferences, using Bayes rule in 8, repeated in 27 below.

$$(27) \quad P_L(w|u) = \frac{P_S(u|w) \times P(w)}{\sum_{w'} P_S(u|w') \times P(w')}$$

We have obtained the empirical estimates for the two terms $P(w)$ and $P_S(u|w)$ in Experiment 3 and Experiment 4. For convenience, we first summarize the results from these two experiments in Table 5 and 6, for the positive and negative predicates separately, and then compute the pragmatic listener's posterior probabilities.

< INSERT TABLE 5 ABOUT HERE >

(28)

$$P_L(w_1|u_{positive}) = \frac{P_S(u|w_1) \times P_L(w_1)}{\sum_{w'} P_S(u|w') \times P_L(w')}$$
$$= \frac{0.48 \times 0.53}{0.48 \times 0.53 + 0.31 \times 0.47}$$
$$= 0.64$$
$$P_L(w_2|u_{positive}) = \frac{P_S(u|w_2) \times P_L(w_2)}{\sum_{w'} P_S(u|w') \times P_L(w')}$$
$$= \frac{0.31 \times 0.47}{0.48 \times 0.53 + 0.31 \times 0.47}$$
$$= 0.36$$

Upon hearing the target utterance, the pragmatic listener's posterior probability for w_1 is higher (0.64) than w_2 (0.36). Recall that in the Truth Value Judgment task (Experiment 1, example 7, the context presented to the participants was the following:
Context for the Truth Value Judgment task: At a recent archaeology conference, Emily said that her research team found evidence to prove that a famous ancient city was actually built by aliens. But she kept the name of the city a secret. (在最近的一次考古界的学术会议上, 艾米丽说她的团队找到了证据证实某一个有名的古城市其实是外星人建造的。但目前她对这个城市的名字保密.)
Because the state w_1 is contradicting what the context scenario describes, a higher posterior probability for w_1 predicts that participants should have been more likely to answer *False* when they were asked in Experiment 1 whether the target sentence fit the given context. Indeed, participants gave more *False* responses when they were presented with 7a in Experiment 1: the predicted posterior probability for w_1 was 0.64, and the mean proportion of responding *False* for 7a in Experiment 1 was 0.73.

For utterances containing a negative predicate, such as 7b, the calculation process is very similar, as shown below.

< INSERT TABLE 6 ABOUT HERE >

(29)

$$P_L(w_1|u_{negative}) = \frac{P_S(u|w_1) \times P_L(w_1)}{\sum_{w'} P_S(u|w') \times P_L(w')}$$
$$= \frac{0.44 \times 0.42}{0.44 \times 0.42 + 0.36 \times 0.58}$$
$$= 0.47$$

$$P_L(w_2|u_{negative}) = \frac{P_S(u|w_2) \times P_L(w_2)}{\sum_{w'} P_S(u|w') \times P_L(w')}$$
$$= \frac{0.36 \times 0.58}{0.44 \times 0.42 + 0.36 \times 0.58}$$
$$= 0.53$$

Upon hearing the target utterance 7b, the pragmatic listener's posterior probability for w_1 is lower (0.47) than w_2 (0.53). Because the w_2 state is consistent with the context scenario provided in the truth value judgment task, a higher posterior probability for w_2 predicts that participants should be more likely to answer *True* in Experiment 1. This again correctly derives the results in Experiment 1, that participants had a higher proportion of *True* responses when they were presented with 7b. There is a potential discrepancy though. The posterior probability for w_2 is 0.53. This predicts a moderate preference of the *True* response for 7b in Experiment 1, but the actual mean proportion of the *True* responses for 7b was 0.73. This mismatch may stem from additional factors that influence truth value judgments. We will discuss one possible source of influence – from questions under discussion (QUDs) – in the General Discussion section.

To summarize, in section 4 we showed that the empirical production results quite closely match the model predicted production patterns in section 3.5. Furthermore, the empirically derived pragmatic listeners' inferences, calculated using the empirical production and prior data, also match reasonably well the overall patterns of the empirical truth value judgment results obtained in Experiment 1. Both findings further corroborate the general proposal laid out in section 3.

5. General Discussion. There are two major findings in this paper. The first finding is an empirical one. Specifically, Experiment 1 and 2 identified an interesting paradox. For a scope-ambiguous wh-in-situ construction in Mandarin, the parser prefers the local scope

dependency, consistent with previous known parsing strategies recruited for dealing with many other types of long distance dependencies. The interpretive bias, however, points in the opposite direction: the interpretation compatible with the high scope dependency is the dominant interpretation. The pursuit of an explanation for this paradox led to the second main finding of this paper: A bayesian pragmatic model, built upon the rational speech act framework (Frank & Goodman 2012, Goodman & Frank 2016), could provide a principled (at least partial) explanation of the interpretation bias, while also incorporating the parsing bias into the same model. In this section, we discuss the general implications of the current proposal and also some limitations.

5.1. Bridging the gap between parsing and interpretation. The experimental findings in the current study contribute new empirical evidence to the observation that there could be misalignments between parsing and interpretation. As mentioned in the Introduction section, similar misalignments have been found in previous work showing that comprehenders can derive interpretations incompatible with the grammatically licensed parse. Almost all existing approaches for addressing this issue focus on rethinking how parsing works. The good enough model (Christianson et al. 2001, Ferreira & Patson 2007) proposes that the parsing outcome may not be a single complete parse, and interpretations can be derived through heuristics instead of a fully specified parse. Some evidence, however, suggests that comprehenders do not necessarily underspecify syntactic structures even when they misinterpret a sentence (Slattery et al. 2013). The noisy channel account (Levy 2008, Gibson et al. 2013) hypothesizes that there is uncertainty in the linguistic input a comprehender perceives, and this introduces distorted alternatives as potential candidates for parsing. The self-organizing model (Tabor & Hutchins 2004) allows a set of lexically anchored tree fragments to form a network via spreading activation, making it possible for locally coherent but globally ungrammatical parses to survive, which in turn explains why people sometimes accept interpretations that are not supported by the global parse (Konieczny et al. 2009). The approach we put forward in this paper departs from these previous approaches by rethinking instead the mapping between parsing outcome and interpretation. Our proposal is grounded in the idea that a comprehender's task is not only to structurally represent the heard utterance, but also (or even more importantly) to infer a message or the communicative intent from the utterance. We maintain the basic parsing assumption that the original linguistic input is fully parsed into grammatical structures, but we introduce pragmatic reasoning to operate on the parsing outcome in order to derive the ultimate interpretation[13].

The current proposal adopts the rational speech act framework. The original RSA

model, focusing primarily on accounting for pragmatic phenomena such as scalar implicatures, deals with syntactically simple and unambiguous utterances (e.g. see a review in Goodman & Frank 2016). In the current study, we extend the original model to syntactically complex domains. We assume a parallel parser that maintains multiple possible parses of a sentence, with parsing biases represented as a probability distribution over all the possible parses. But there is not a simple correspondence between parsing biases and the interpretations obtained by a listener. The linguistic update of a listener is determined by the interaction between parsing biases and a number of other factors. To start with, in order to incorporate parsing biases into the pragmatic reasoning process, we combine the effect of different parses based on the probability of each parse when calculating the linguistic update at the *literal listener* stage. As shown by the examples in 15, 17, and 19, three factors work together to determine a literal listener's linguistic update: parsing biases, the prior probabilities of the relevant world states, and the compatibility between a world state and a particular parse of the utterance. This means that parsing decisions alone do not necessarily determine a literal listener's linguistic update. While one parse may be compatible with only one relevant world state, another parse may be compatible with more than one world state. Different world states also have different prior probabilities. Due to the interaction between these different factors, even when the parser strongly favors a particular parse, it is still possible that the interpretation (or the world state) supported by that parse does not become the dominant one for the literal listener. Conversely, a world state compatible with a dispreferred parse still has the chance to become a strong candidate in the posterior beliefs of a literal listener. This is a key feature of the current proposal. It affords a more flexible mapping between parsing and interpretation in a principled manner, allowing potential misalignment between parsing preferences and interpretive preferences from the very beginning of the recursive pragmatic reasoning process. The effect of the parsing bias, entering the pragmatic reasoning process at the literal listener stage, will eventually percolate up and have an influence on the linguistic update of the pragmatic listener, mediated by the intermediate pragmatic speaker. In addition, the linguistic update of the pragmatic listener is also affected by the prior probability of each relevant world state (see the calculations in 28 and 29. Taking everything together, parsing bias does not directly determine interpretation; instead, it becomes part of the overall pragmatic reasoning process that gives rise to the ultimate interpretation. Once we remove the premise that there is a direct mapping between parsing outcome and interpretation, the seeming paradox we observed earlier between parsing and interpretation in effect disappears.

In addition to accounting for the seeming paradox presented in the current case study,

i.e. the interpretation favored by the dominant parse is not chosen as the dominant interpretation, the current proposal also predicts that an interpretation supported by a high prior probability is also not necessarily the winning candidate. Again this is because the bayesian update of a pragmatic listener is conditioned by a number of factors together, as discussed above, instead of any single factor alone. This prediction is in line with some examples raised by a reviewer, such as *The workers painted the doors with cracks* or *The girl sliced the apple with a blemish*. In these examples, the implausible interpretation (e.g. "cracks" or "blemish" was interpreted as an instrument argument) is the more dominant interpretation even though the world states they represent should have very low prior probabilities. Based on the current proposal, we speculate that these examples are likely to have a strong parsing bias that attaches the prepositional phrase as a verbal adjunct, and when the parsing bias and the speaker's production probability are taken into account, the pragmatic listener's posterior probability will turn out to favor the world state that has very low priors.

Since the current case study involves structural ambiguity (i.e. the high vs. low scope of the in-situ wh-phrase), it is worth noting that in the syntactic ambiguity resolution literature, there was an influential debate about how syntactic information and other sources of information should be integrated. The central empirical domain of this body of work largely focuses on garden-path ambiguity resolution. Consider the following garden-path ambiguity example. The partial sentence "The witness examined..." could be interpreted as denoting either a subject-predicate relation or a subject-modification (with a reduced relative clause) relation depending on whether the verb "examined" is parsed as a matrix verb or a past participle. These two parses could receive differential support from syntactic, lexical and contextual/pragmatic information. For instance, parsing "examined" as a matrix verb may be the temporarily preferred parse since it is structurally less complex than the alternative parse that postulates a reduced relative clause; but on the other hand, the interpretation of the relative clause parse could be pragmatically more felicitous depending on the context. The competition and trade-off between different sources of information also creates a kind of "misalignment" – in this case, the parse favored for structural complexity reasons could be in conflict with the parse supported by pragmatic context. When broadly construed, the question of how to resolve this sort of conflicts can also be viewed as addressing a related problem as the current study. But we note that the theoretical focus is not the same. The large body of work on garden-path ambiguity resolution aims at understanding how the parser combines multiple sources of information to guide its parsing decision. The various answers to the question range from proposals that prioritize structural principles to guide the initial

parsing decision, while consulting other sources of information in the later structural reanalysis process (e.g. Frazier 1978, Frazier & Fodor 1978), to proposals that view parsing as a constraint-satisfaction process, that integrates all sources of relevant information as quickly as possible to arrive at the correct parse (MacDonald et al. 1994, McRae et al. 1998, Trueswell et al. 1994). The current case study, although involving structural ambiguity, is not concerned with how people consult different sources of information to choose between a high-scope versus a low-scope parse. Instead, the empirical puzzle is that even after establishing the fact that people have settled on the low-scope parse as the preferred parse at the global level, the interpretation compatible with the low-scope parse is still not perceived as the preferred interpretation of the utterance. It is this kind of misalignment between the parsing outcome and the ultimate interpretation at the global level that we aim to account for.

One may ask if it is possible that the initially preferred low-scope parse was somehow reanalyzed into a high-scope parse in the truth value judgment experiment. Reanalysis is possible for the classic garden-path sentences due to the presence of clear error signals and disambiguating cues. But we are not aware of any systematic cues in our experiment that would trigger a reanalysis on the scope dependency. With that said, although our theoretical goal is not entirely identical with the syntactic ambiguity resolution literature on garden-path sentences, future work should still explore whether the current proposal could be extended to shed new light on a broader range of phenomena regarding parsing and interpretation, including the garden-path ambiguity resolution problem. The proposal we outlined here only integrates parsing and pragmatic reasoning at the global utterance level. If the general approach could be extended to incrementally integrate parsing and pragmatic reasoning for partial utterances (see Cohn-Gordon et al. 2019 for a proposal of an incremental RSA model), this may provide a new way to model a number of classic problems of incremental comprehension. For example, as pointed out by a reviewer, when integrating syntactic and (non-syntactic) contextual information to resolve temporary garden-path ambiguity, the most common method in the literature is to implement a probabilistic cue-weighting strategy (e.g. Narayanan & Jurafsky 1998, McRae et al. 1998), i.e. different sources of information are combined by a weighting parameter that determines how strong an effect each type of information bears upon the ultimate parsing choice and interpretation. Determining the value of the weighting parameter in a principled manner, however, could be theoretically challenging. In the current proposal, integrating parsing and pragmatic reasoning does not evoke cue-weighting. Instead, parsing biases are fully embedded within the bayesian pragmatic reasoning process. This feature is potentially theoretically appealing, and it could open up new possibilities to

account for incremental comprehension.

5.2. The potential role of QUDs. Although the bayesian pragmatic model provided good qualitative predictions for the interpretation bias, as we noted earlier, it did not completely capture the behavioral results from the truth value judgment task. The mismatch was more salient when the utterance contained a negative matrix predicate – for an utterance like 7b, the model only predicted a moderate bias for the *true* response (53% posterior probability for the w_2 state that will lead to a *true* response, see 29, whereas the empirical results in Experiment 1 showed a more substantial bias for the *true* response (73%). This discrepancy suggests to us the current analysis needs further refinement. We speculate here that making the analysis more sensitive to the relevant questions under discussion (QUD, Ginzburg 1996, Roberts 1996) could potentially lead to improvement. A structured discourse can be perceived as being organized around a set of *issues* or *questions* that the interlocutors are committed to resolving together. Each sentence coheres with the previous discourse context by virtue of helping to address the currently shared (often implicit) QUD at that given moment in time, for instance, by providing an answer to it or by raising another relevant question. A comprehender could approach a given utterance as an answer to a discourse-salient QUD, and her pragmatic inference should be conditioned by this currently relevant QUD. A number of previous studies have explored how to incorporate QUDs into the RSA models (Degen & Goodman, 2014, Savinelli et al. 2018, Scontras & Goodman 2017). One empirical challenge with this approach is that there is no currently known rigorous method to systematically track (implicit) QUDs in a discourse context[14]. With this caveat in mind, we sketch a suggestion below that could potentially better model the truth value judgment results by introducing QUDs into the current proposal.

With the truth value judgment task, recall that in our working example 7, the context scenario ended with a note that Emily kept secret the name of the city in her team's discovery. This last sentence may have made the naming event highly salient for at least some participants. These participants could be motivated to construct an implicit QUD like "Did Emily announce the name of the city?". When they then received a target sentence and was asked to judge whether the target sentence fits the context scenario, they may have based their *true/false* judgments largely on how the target sentence answers this QUD and whether that answer is congruent with the context. In 28 and 29 we have computed the participants' posterior probabilities of different world states after receiving a positive or negative target utterance. It is crucial to note that for an utterance containing a positive predicate, the two relevant world states in Table 5 would

provide different answers to the QUD "Did Emily announce the name of the city?". The w_1 state is a world state that will trigger the answer "*Yes, she did*" to the implicit QUD. This answer contradicts how the QUD was actually resolved in the context scenario, and therefore a comprehender that endorses w_1 would judge that the target sentence *does not fit* or *false* under the given context. The w_2 state, on the other hand, will trigger the answer "*No, she didn't*" to the implicit QUD, consistent with how the QUD was resolved in the context scenario, leading to a truth value judgment *fits* or *true*. We predict that the *true/false* responses in Experiment 1 should track the posterior probabilities of the w_1/w_2 states, which by and large was indeed the case. But the situation for target utterances containing a negative predicate is a little different. The two relevant world states in Table 6 would both trigger the same answer "*No, she didn't*" to the implicit QUD, regardless of the listener's posterior preferences for these two world states. This would mean that a participant should always conclude that the target sentence answered the QUD in a way consistent with how the QUD was resolved in the context. Therefore the target sentence has a very high probability to be judged as *fits* or *true* under the context. This could explain why in Experiment 1, the proportion of responding *true* for sentence containing a negative predicate is much higher than the model predicted posterior probability for the w_2 state in 29.

Under the scenario outlined above, the basic belief update process remains the same as our original proposal, and participants' sensitivity to QUDs only has an effect at the last step of completing the truth value judgment task: rather than directly evaluating whether each updated world state is consistent with the context scenario, participants instead evaluate whether each updated world state answers the discourse salient QUD in a way consistent with how the QUD is resolved in the context. Alternatively, it is also possible that QUDs could make contributions at a much earlier stage. For instance, primed by the implicit QUD "Did Emily announce the name of the city?", participants may decide to prioritize the parse that could clearly answer the QUD. Since only the high scope parse clearly specifies (the low scope parse underspecifies) whether the naming event happened, participants may be led to favor the high scope parse and give their truth value judgments based on the high scope parse. In this way, QUDs play a role in actually shaping participants' early parsing decisions. The hypothesis that QUDs can have an effect on early parsing decisions finds independent support from some previous evidence (e.g. Clifton & Frazier 2012), but there is a potential challenge for this hypothesis when the acceptability results from Experiment 2 are considered. Experiment 2 shares identical context scenarios and target sentences with Experiment 1. If contextually triggered implicit QUDs can guide participants to more readily settle on a high scope

parse, this may incorrectly predict that participants could have overcome the locality bias in Experiment 2 and given higher acceptability ratings for sentences that only have a high scope parse (i.e. the unambiguous conditions).

Lastly, our discussion about QUDs so far still assumes an idealized listener who can build complete parses and integrate the parsing outcome with the pragmatic reasoning process. There is yet another possibility. With complex sentences like the ones we tested here, participants may develop strategies to answer the implicit QUD "Did Emily announce the name of the city?" without fully parsing the target sentence. For example, they may have simply remembered the beginning of the target sentence "Emily announced" or "Emily concealed", and used those sentence fragments to answer the QUD, and then derived the truth value judgments by comparing whether the target sentences answered the QUD in the same way as how the QUD was resolved under the context scenario in the experiment. We can not rule out this possibility. But we note that although this possibility may seem simpler than what we outlined above, the simplicity comes with a theoretical disadvantage. Since this "partial-sentence" heuristic only narrowly targets the truth value judgment task, it would be completely silent on how to account for the production preferences observed in Experiment 4, and an account of the latter has to be separately stipulated. The proposal we developed offers a more principled way to cover a broader range of empirical findings.

5.3. Analysis at the individual item level. The analysis we presented in section 4.2 was based on data aggregated over participants and items. It is worth asking whether the pragmatic model we used to explain the truth value judgments at the population level can also explain individual behavior. Unfortunately, as the truth value judgments, the prior estimates, and the production bias estimates in the current study were collected from different groups of participants, we are not able to construct a pragmatic model for each participant. But as a proof of concept, we nonetheless carried out a by-item analysis and examined whether the bayesian pragmatic reasoning introduced in section 5 could explain, at least to some extent, the truth value judgments obtained for each item.

Recall that in the current study we constructed 16 sets of scenarios/items like the ones presented in example 7. The same set of material, modified for the specific goals of different experiments, were used to collect truth value judgments, prior estimate and production bias estimates. We therefore could do the calculation in 28 and 29 for each item separately, and then correlate, at the individual item level, the posterior probability obtained for a world state and the truth value response consistent with that world state. Due to an experiment error, one item used to estimate the production bias in Experiment

4 had a slightly different predicate from the same item used in the other experiments. We excluded this item from the by-item correlation analysis. The correlation results obtained from 15 items are plotted in Figure 6.

< INSERT FIGURE 6 ABOUT HERE >

In Figure 6, for each target sentence with a positive predicate (Figure 6, Left), we correlated the proportion of *false* (e.g. *does not fit*) responses with the posterior probabilities of the w_1 state. The y-axis in the figure represents the proportion of the *false* responses for each item. *False* is the majority response obtained for the positive predicates in Experiment 1. Since in Experiment 1 the w_1 state supports the *false* judgment for a positive predicate, the x-axis in the figure represents the posterior probability of the w_1 state.[15] The calculation for the posterior probability is identical to the calculation of $P_L(w_1|u_{positive})$ in 28, except that it is now done for each individual item. Similarly, for each target sentence with a negative predicate (Figure 6, Right), we correlated the proportion of the *true* (e.g. *fit*) responses to the posterior probability of the w_2 state. *True* is the majority response obtained for the negative predicates in Experiment 1, and the w_2 state is the world state that supports the *true* judgment. A significant correlation would indicate that, at the individual item level, the posterior probabilities of the relevant world states derived by the pragmatic model are indeed related to the experimentally estimated truth value judgments. But as shown in Figure 6, there are no significant correlations (ps>.3).

One possible source for the lack of correlation in the by-item analysis is the estimated prior probabilities for each world state. The scenarios we constructed for the current study are all somewhat arbitrary; at the individual scenario level, the prior probability estimate for each world state may have been too noisy. We did an exploratory analysis that removed the effect of the prior from the calculation. This amounts to assuming an equal prior probability for the two alternative world states at the individual scenario level, with $P(w_1) = P(w_2) = 0.5$. The by-item correlation under this new analysis is presented in Figure 7. In Figure 7, for the positive predicate, the y-axis still represents the truth value judgments consistent with the world state w_1 (i.e. the proportion of *false* judgments); the x-axis, instead of representing the posterior probability of w_1, represents how likely a speaker is to choose the target wh-in-situ structure to describe w_1, given that the speaker could use the target structure to describe either w_1 or w_2[16]. For the negative predicate in Figure 7, the y-axis represents the truth value judgments consistent with the world state w_2 (i.e. the proportion of *true* judgments), and the x-axis represents how likely a speaker is to choose the target wh-in-situ structure to describe w_2.

< INSERT FIGURE 7 ABOUT HERE >

The by-item correlation is marginal for the positive predicate items ($p<.06$), and significant for the negative predicate items ($p<.05$). This exploratory analysis provides some very preliminary evidence that at the individual item level, a listener's truth value judgment is correlated with the production choice of the speaker. Overall, however, there is no strong conclusion we can draw at the individual item level. More future work is needed to address questions about individual variations.

6. Conclusion. To conclude, focusing on the wh-in-situ scope ambiguity in Mandarin Chinese, our study provides novel empirical evidence to show that parsing and interpretation decisions at the global level can misalign. We develop an analysis that incorporates parsing decisions into a general bayesian pragmatic reasoning architecture, circumventing any actual conflict between parsing and interpretation. Our study therefore brings closer two strands of research in psycholinguistics, one on structure parsing, and the other on pragmatic reasoning.

REFERENCES

Alexopoulou, Theodora, and Frank Keller. 2007. Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language*. 110-160.

Aoun, Joseph, and Yen-hui Audrey Li. 1993. Wh-elements in situ: Syntax or LF? *Linguistic Inquiry*. 24. 199-238.

Cheng, Lisa Lai Shen. 1991. *On the typology of wh-questions*. PhD dissertation, MIT.

Chomsky, Noam, and George A Miller. 1963. Introduction to the formal analysis of natural languages. In Luce, R. Duncan, Robert R. Bush, and Eugene Galanter (Eds), *Handbook of mathematical psychology*. Vol 2. 269-321.

Christianson, Kiel; Andrew Hollingworth; John F Halliwell; and Fernanda Ferreira. 2001. Thematic roles assigned along the garden path linger. *Cognitive Psychology*. 42. 368–407.

Clifton Jr, Charles, and Lyn Frazier. 2012. Discourse integration guided by the 'question under discussion'. *Cognitive Psychology*. 65(2). 352-379.

Cohn-Gordon, Reuben; Noah D Goodman; and Christopher Potts. 2018. An incremental iterated response model of pragmatics. *Proceedings of the Society for Computation in Linguistics (SCiL)*. 81-90

Cuetos, Fernando and Don C Mitchell. 1988. Cross-linguistic differences in parsing: Restrictions on the use of the Late Closure strategy in Spanish *Cognition*. 30(1). 73-105

Degen, Judith; and Noah Goodman. 2014. Lost your marbles? The puzzle of dependent measures in experimental pragmatics. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. 397-402.

Drummond, Alex. 2016. Ibex Farm. https://github.com/addrummond/ibex. GitHub.

Fanselow, Gisbert, and Stefan Frisch. 2006. Effects of processing difficulty on judgments of acceptability. *Gradience in grammar: Generative perspectives*. 291-316.

Ferreira, Fernanda; Karl GD Bailey; and Vittoria Ferraro. 2002. Good-enough representations in language comprehension. *Current Directions in Psychological Science*. 11(1). 11–15.

Ferreira, Fernanda; Kiel Christianson; and Andrew Hollingworth. 2001. Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*. 30(1). 3–20.

Ferreira, Fernanda, and Nikole D Patson. 2007. The "good enough" approach to language comprehension. *Language and Linguistics Compass*. 1. 71–83.

Ferreira, Fernanda. 2003. The misinterpretation of noncanonical sentences. *Cognitive psychology*. 47(2). 164-203

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

Frazier, Lyn, and Janet Dean Fodor. 1978. The sausage machine: A new two-stage parsing model. *Cognition*. 6(4). 291-325.

Gibson, Edward. 1998. Linguistic complexity: Locality of syntactic dependencies. *Cognition*. 68. 1-78.

Gibson, Edward. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*. 95-126.

Gibson, Edward; Leon Bergen; and Steven T Piantadosi. 2013. Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*. 110. 8051–8056.

Ginzburg, Jonathan. 1996. Dynamics and the semantics of dialogue. *Logic, language and computation*. 1. 221-237.

Goodman, Noah D, and Michael C Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*. 20(11). 818-829.

Grice, Herbert P. 1975. Logic and conversation. In Peter Cole and Jerry Morgan (eds.), *Syntax and semantics*. 3. *Speech Acts*. 41–58. New York: Academic Press.

Hale, John. 2003. The information conveyed by words in sentences. *Journal of Psycholinguistic Research*. 32(2). 101-123.

Huang, C-T James. 1982. *Logical relations in Chinese and the theory of grammar*. Ph.D. dissertation, MIT.

Hofmeister, Philip; T Florian Jaeger; Inbal Arnon; Ivan A Sag; and Neal Snider. 2013. The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes* 28. 48-87.

Konieczny, Lars; Daniel Müller; Wibke Hachmann; Sarah Schwarzkopf; and Sascha Wolfer. 2009. Local syntactic coherence interpretation. Evidence from a visual world study. *Proceedings of the 31st annual conference of the Cognitive Science Society*. 1133-1138.

Levy, Roger. 2008a. Expectation-based syntactic comprehension. *Cognition*. 106(3). 1126-1177.

Levy, Roger. 2008b. A noisy-channel model of rational human sentence comprehension under uncertain input. *Proceedings of the 13th Conference on Empirical Methods in Natural Language Processing*. 234–243. Association for Computational Linguistics, Stroudsburg, PA.

Lewis, Richard L; Shravan Vasishth; and Julie A Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in cognitive sciences*. 10(10). 447-454.

Lewis, Richard L, and Shravan Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*. 29. 375-419.

MacDonald, Maryellen C; Neal J Pearlmutter; and Mark S Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological review*. 101(4), 676-703.

McRae, Ken; Michael J Spivey-Knowlton; and Michael K Tanenhaus. 1998. Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*. 38(3). 283-312.

Narayanan, Srini, and Daniel Jurafsky. 1998. Bayesian models of human sentence processing. *Proceedings of the twelfth annual meeting of the cognitive science society*. 752-757

Roberts, Craige. 1996. Information structure in discourse: Toward an integrated for-mal theory of pragmatics. *Ohio State University Working Papers in Linguistics*. 49. 91-136.

Ronai, Eszter, and Ming Xiang. 2021. Pragmatic inferences are QUD-sensitive: an experimental study. *Journal of Linguistics.* 57(4). 841-870.

Qian, Zhiying; Susan Garnsey; and Kiel Christianson. 2018. A comparison of online and offline measures of good-enough processing in garden-path sentences. *Language, Cognition and Neuroscience*. 33(2). 227-254.

Savinelli, KJ; Greg Scontras; and Lisa Pearl. 2018. Exactly two things to learn from modeling scope ambiguity resolution: Developmental continuity and numeral semantics. *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics*. 67-75.

Scontras, Gregory, and Noah D Goodman. 2017. Resolving uncertainty in plural predication. *Cognition*. 168. 294-311.

Slattery, Timothy J; Patrick Sturt; Kiel Christianson; Masaya Yoshida; and Fernanda Ferreira. 2013. Lingering misinterpretations of garden path sentences arise from competing syntactic representations. *Journal of Memory and Language*. 69(2). 104-120.

Tabor, Whitney; Bruno Galantucci; and Daniel Richardson. 2004. Effects of merely local syntactic coherence on sentence processing. *Journal of Memory and Language*. 50(4). 355-370.

Tabor, Whitney, and Sean Hutchins. 2004. Evidence for self-organized sentence processing: digging-in effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. 30(2). 431-450.

Tanenhaus, Michael K; Michael J Spivey-Knowlton; Kathleen M Eberhard; and Julie C Sedivy 1995. Integration of visual and linguistic information in spoken language comprehension. *Science*. 268(5217). 1632-1634.

Traxler, Matthew J; Martin J Pickering; and Charles Clifton Jr. 1998. Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*. 39. 558-592.

Trueswell, John C; Michael K Tanenhaus; and Susan M Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of memory and language*. 33(3). 285-318.

Tsai, Wei-Tien Dylan. 1994. *On economizing the theory of A-bar dependencies*. Ph.D. dissertation, MIT.

Van Dyke, Julie A, and Richard L Lewis. 2003. Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language*. 49. 285-316.

Van Gompel, Roger PG; Martin J Pickering; and Matthew J Traxler. 2000 Unrestricted race: A new model of syntactic ambiguity resolution *Reading as a perceptual process*. 621-648. Elsevier.

Warren, Tessa, and Edward Gibson. 2002. The influence of referential processing on sentence complexity. *Cognition*. 85. 79-112.

Xiang, Ming; SuiPing Wang; and YanLing Cui. 2015. Constructing covert dependencies-The case of Mandarin wh-in-situ dependency. *Journal of Memory and Language*. 84. 139-166.

Xiang, Ming, and SuiPing Wang. 2020. Locality and Expectation in Chinese wh-dependencies. University of Chicago, MS.

NOTES

[1]Based on the binary yes-no acceptability judgments reported in Xiang and Wang (2020, Experiment 2), sentences like 3a and 3b were rated on average at 0.7 and 0.3, and sentences like 5a and 5b were rated on average at 0.71 and 0.67.

[2]We chose to use the wording *(does not) match* instead of *true/false* because the literal translation of latter in Mandarin sounded unnatural as task instructions.

[3]The converged mixed effects logistic model is: model=$glmer(acceptability\sim ambiguity\times verbpolarity + (1 + ambiguity|subj) + (1|item), data, family = binomial)$

[4]The experiments in Xiang and Wang (2020) did not manipulate the polarity of the matrix verb, in fact, the majority of the items there have a positive matrix predicate. The basic sentence structures tested there are identical to the target sentences in 7, but they do not have a preceding context scenario. The mean acceptability judgments (yes/no binary judgments) reported in Experiment 1 of Xiang and Wang (2020) were 0.75 for the ambiguous condition, and 0.39 for the unambiguous condition; these ratings were replicated in their Experiment 2, with 0.7 for the ambiguous condition and 0.3 for the unambiguous condition. There were also completely ungrammatical filler sentences included in their study, which received an average acceptability rating of 0.18. The mean ratings for the ambiguous and unambiguous conditions from Xiang and Wang (2020) are on a par with the mean ratings in the current study, and the slightly lower ratings on the ambiguous positive predicate sentences in the current study are more likely due to the context scenario instead of the sentence complexity itself.

[5]We make the assumption that the prior is shared for both the pragmatic listener inference in 8 and the literal listener inference in 11. Therefore there is no subscript on the term $P(w)$.

[6]Strictly speaking, the utility function should also consider the cost of an utterance. For the sake of simplicity, we only consider the informativeness of an utterance here.

[7]The derivation in 12 holds because $P(u) = P(u_h) + P(u_l) = 1$, assuming that the current target utterance u only has two parses u_h and u_l. The full derivation is the following: $P_{L_0}(w|u) = \frac{P(w \cap u)}{P(u)} = P_{L_0}(w \cap u_h) + P_{L_0}(w \cap u_l)$, since $u = u_h \cup u_l$, $u_h \cap u_l = \emptyset$, and $P(u) = 1$. Furthermore, $P_{L_0}(w \cap u_h) = P_{L_0}(w|u_h) \times P(u_h)$, and $P_{L_0}(w \cap u_l) = P_{L_0}(w|u_l) \times P(u_l)$.

[8]The goal of the prior elicitation task is to establish priors for the world states relevant to the interpretation of the target utterance. One potential concern could be that the context in 14 should be biased instead of neutral, in keeping with the context used for the truth value judgment task in experiment 1. But in our truth value judgment task, the target sentence was meant to be an independent sentence, instead of a continuation of the context sentence. As shown in the procedure of Experiment 1, the context scenario and the target sentence were presented on two separate screens, and participants were explicitly asked to decide whether the meaning of the target sentence matches or did not match the context scenario they saw. For participants to give a truth value judgment, they need to compare their interpretation of the target sentence to the (biased) context, but this does not mean they condition the interpretation of the target utterance on the context. In fact, if they had conditioned the interpretation of the target sentence on the context, they should have always interpreted the target sentence in a way that was consistent with the context, and should have not ever given a *false* response. We therefore did not use the biased context from Experiment 1 for the prior elicitation task.

[9]The complexity of the target constructions in the current study also makes it more difficulty to estimate production cost. As alluded to in section 3.1, a speaker's choice between alternative utterances should in principle reflect a trade-off between the informativity of an utterance and the cost of that utterance. But even for simple utterances, there is no currently known satisfying metric to precisely quantify utterance cost. The problem is further complicated by complex syntactic structures like the ones investigated in this paper, since many aspects of a complex sentence could contribute to production cost, such as sentence length, syntactic complexity, information structure, ambiguity resolution, etc. For simplicity, we did not consider production cost in the current study.

[10]When $0 < y < x < 1$ and $\alpha > 0$, $x^\alpha y^\alpha + x^\alpha > x^\alpha y^\alpha + y^\alpha$, and it follows from there that $\frac{x^\alpha}{x^\alpha + 1} > \frac{y^\alpha}{y^\alpha + 1}$.

[11]The converged mixed effects model is: model = $glmer(response{\sim}VerbPolarity \times worldstate + (1|subj) + (1 + worldstate|item), data = data, family = binomial)$. Both predictors are sum-coded.

[12]The English translation may look like passivization, but the actual Mandarin production often involves a focus marker "shi" to front the wh-phrase to the clause initial position.

[13]We also note that the proposal presented here, although incorporating the parsing outcome into the pragmatic reasoning process, does not aim to explain what leads to the parsing outcome in the first place. Following previous work, we assume there is a set of independent mechanisms that affect the parsing outcome, including the complexity of the to-be-established structure (Frazier & Fodor 1978), working memory constraints (Gibson 1998, Lewis et al. 2006), syntactic or semantic expectation of the upcoming material (Hale 2003, Levy 2008), or contextual influence (Tanenhaus et al. 1995).

[14]A recent study from Ronai and Xiang (2021) did an elicitation experiment to empirically identify potential QUDs pertaining to the calculation of scalar implicatures. In addition to uncovering questions that are consistent with the previous literature, their results also uncovered questions that have not been discussed as relevant to implicature derivation.

[15]Correlating the minority responses from Experiment 1 did not make a difference, e.g. correlating the *true* responses with the posterior probabilities of the $w2$ states for the positive predicates

[16]One could see this more clearly based on the calculation of $P_L(w_1|u_{positive})$ in 28. When $P(w_1)$ and $P(w_2)$ are set to be 0.5 in this equation, the right hand side of the equation is essentially equivalent to $\frac{P_S(u|w_1)}{P_S(u|w_1)+P_S(u|w_2)}$, and this is what the x-axis in Figure 7 (Left) represents. Similarly, for the plot on the right in Figure 7, the x-axis represents $\frac{P_S(u|w_2)}{P_S(u|w_1)+P_S(u|w_2)}$.

A. List of tables.

B. List of figures.

TABLE 1. World states relevant for utterances with positive predicate

world states	e1 name announcement	e2 discovery announcement	Considered as a relevant world state?
w_1	+	+	yes
w_2	+	−	no
w_3	−	+	yes
w_4	−	−	no

TABLE 2. World states relevant for utterances with negative predicate

world states	e1 name concealing	e2 discovery concealing	Considered as a relevant world state?
w_1	+	+	yes
w_2	+	−	yes
w_3	−	+	no
w_4	−	−	no

TABLE 3. A summary of the relevant world states considered in the model

world states	Positive matrix predicate	Negative matrix predicate
w_1	Emily announced they discovered that a city was built by aliens and she also announced the name of the city. (艾米丽宣布了她们发现了有一个城市是外星人建造的，她也同时宣布了这个城市的名字.)	Emily concealed the fact that they discovered that a city was built by aliens and she also concealed the name of the city. (艾米丽隐瞒了她们发现了有一个城市是外星人建造的，她也同时隐瞒了这个城市的名字.)
w_2	Emily announced they discovered that a city was built by aliens but she did not announce the name of the city. (艾米丽宣布了她们发现了有一个城市是外星人建造的，但她没有宣布这个城市的名字.)	Emily did not conceal the fact that they discovered that a city was built by aliens but she conceal the name of the city. (艾米丽没有隐瞒她们发现了有一个城市是外星人建造的，但是她隐瞒了这个城市的名字.)

TABLE 4. Literal listener's posterior probabilities for each pair of utterance and world state

Alternative Utterances	w_1	w_2
u_{ambig}	$P_{L0}(w_1 \mid u_{ambig})$	$P_{L0}(w_2 \mid u_{ambig})$
$u_{unambig1}$	$P_{L0}(w_1 \mid u_{unambig1}) = 1$	$P_{L0}(w_2 \mid u_{unambig1}) = 0$
$u_{unambig2}$	$P_{L0}(w_1 \mid u_{unambig2}) = 0$	$P_{L0}(w_2 \mid u_{unambig2}) = 1$

TABLE 5. For the positive predicate, see the example in 7a and 13a:

Target sentence: Emily *announced* her team *discovered* aliens established which city.
(艾米丽公布了她的团队发现了外星人建造了哪座城市.)

World states	$P_s(u\|w)$	Priors
w_1: Emily announced they discovered that a city was built by aliens, and she also announced the name of the city. (艾米丽公布了她们发现了有一个城市是外星人建造的，她也同时公布了这个城市的名字.)	0.48	0.53
w_2: Emily announced they discovered that a city was built by aliens, but she did not announce the name of the city. (艾米丽公布了她们发现了有一个城市是外星人建造的，但她没有公布这个城市的名字。)	0.31	0.47

TABLE 6. For the negative predicate, see the example in 7b and 13b.

Target sentence: Emily *concealed* her team *discovered* aliens established which city.
(艾米丽隐瞒了她的团队发现了外星人建造了哪座城市.)

World states	$P_s(u\|w)$	Priors
w_1: Emily concealed the fact that they discovered that a city was built by aliens, and also concealed the name of the city.（艾米丽隐瞒了她们发现了有一个城市是外星人建造的，她也同时隐瞒了这个城市的名字。）	0.44	0.42
w_2: Emily did not conceal the fact that they discovered that a city was built by aliens, but she concealed the name of the city. （艾米丽没有隐瞒她们发现了有一个城市是外星人建造的，但是她隐瞒了这个城市的名字。）	0.36	0.58

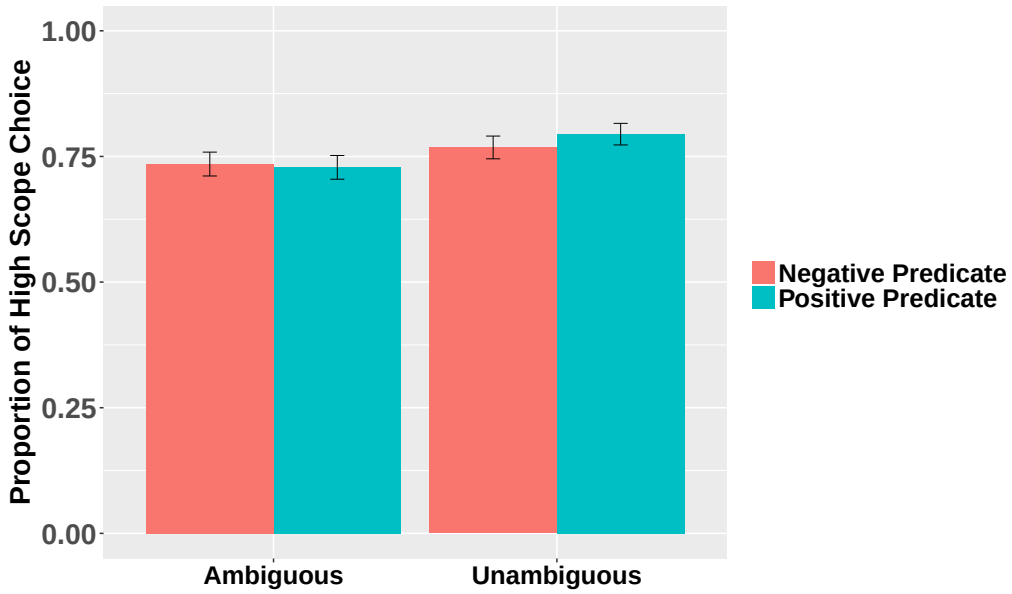


FIGURE 1. Truth value judgment task results: proportion of participants' choosing the high scope construal

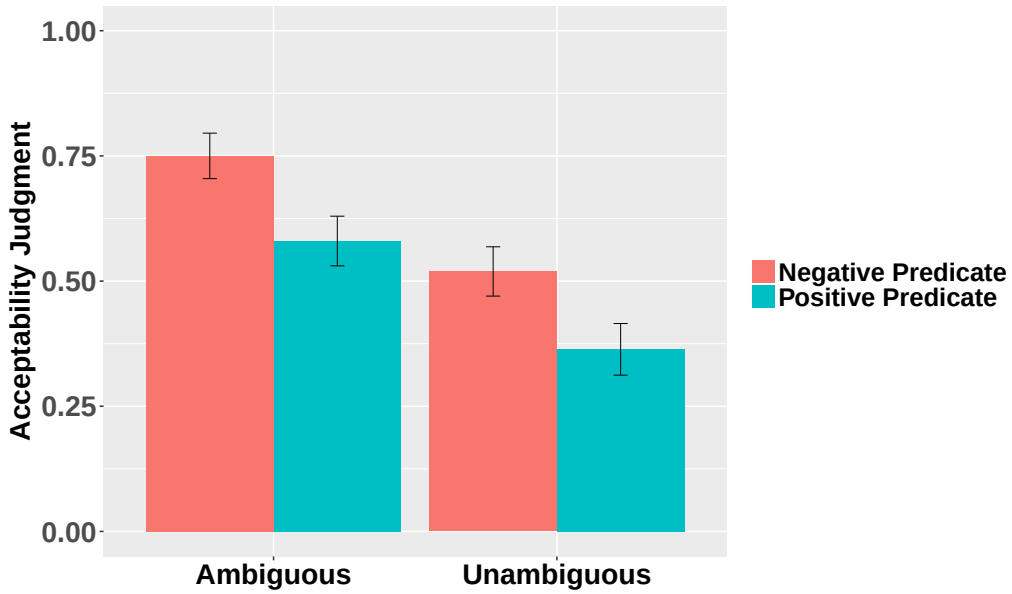


FIGURE 2. Acceptability judgment results: Proportion of participants' Yes responses

At a recent archaeology conference, Emily made a presentation on behalf of her research team. (在最近的一次考古界的学术会议上, 艾米丽代表她的研究团队作了一个报告。)

(On a separate screen)

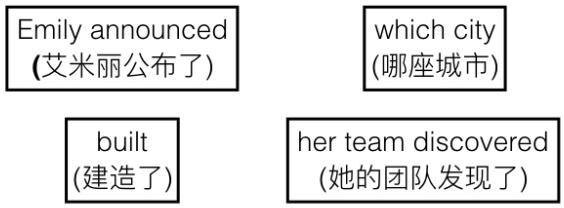
Which of the following situation is more likely to arise? (以下的哪种情况更有可能发生?

1. In her report, Emily said that her research teamfound evidence to prove that a famous ancient city was actually built by aliens. She also released the name of the city. (在她的报告里，艾米丽说她 团队找到了证据证实某一个有名的古城市其实是外星人建造的，她同时也宣布了这个城市的名字。)

2. In her report, Emily said that her research team found evidence to prove that a famous ancient city was actually built by aliens. But the name of the city needs to be kept secret for the moment. (在她的报告里，艾米丽说她的团队找到了证据证实某一个有名的古城市其实是外星人建造的，但目前她需要对这个城市的名字保密。)

FIGURE 3. An example trial for Experiment 3. This example represents a trial that estimates the prior probability of each relevant world state under a positive predicate sentence.

At a recent archaeology conference, Emily said that her research team found evidence to prove that a famous ancient city was actually built by aliens. She also released the name of the city. (在最近的一次考古界的学术会议上, 艾米丽说她的团队找到了证据证实某一个有名的古城市其实是外星人建造的。 她同时还公开了这个城市的名字.)



Please make a sentence based on this scenario. The sentence you make should include the four phrases above, and its content should also be compatible with the scenario. (请根据这个场景造一个句子。您造的句子需要包括以上这四个词，还需要符合场景描述的内容).

FIGURE 4. An example trial for the production experiment reported in Experiment 4

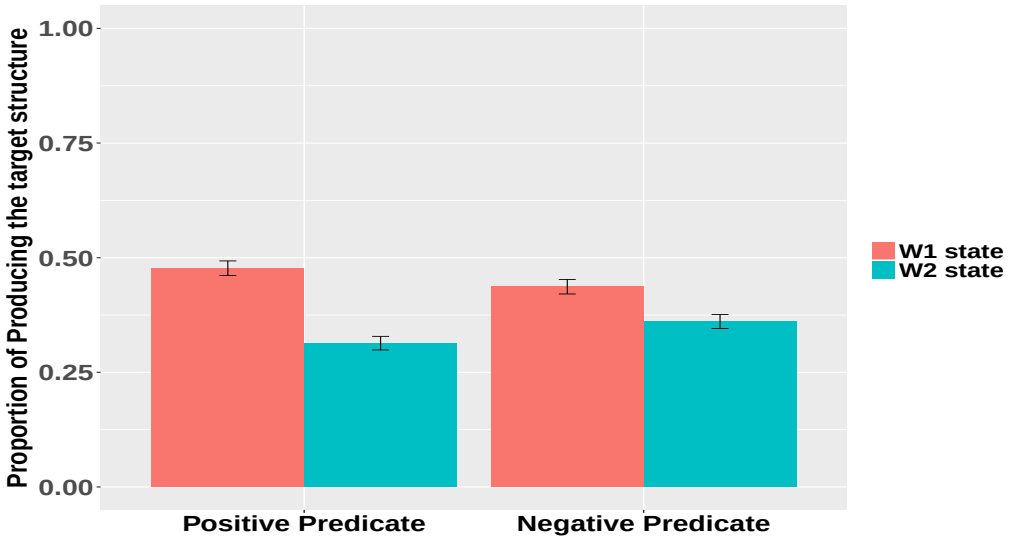


FIGURE 5. Proportion of producing the target wh-in-situ structure

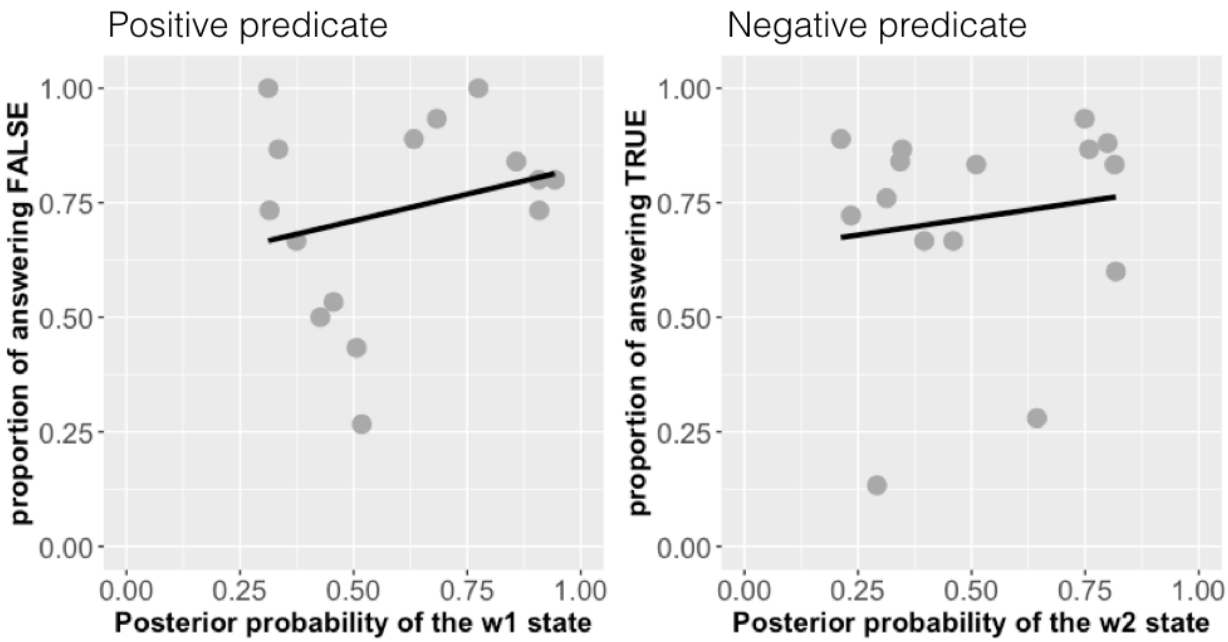


FIGURE 6. By-item correlation between the truth value judgments from Experiment 1 and the posterior probabilities of the relevant world state. Left: positive predicates, $p = 0.37$; Right: negative predicates, $p = 0.6$.

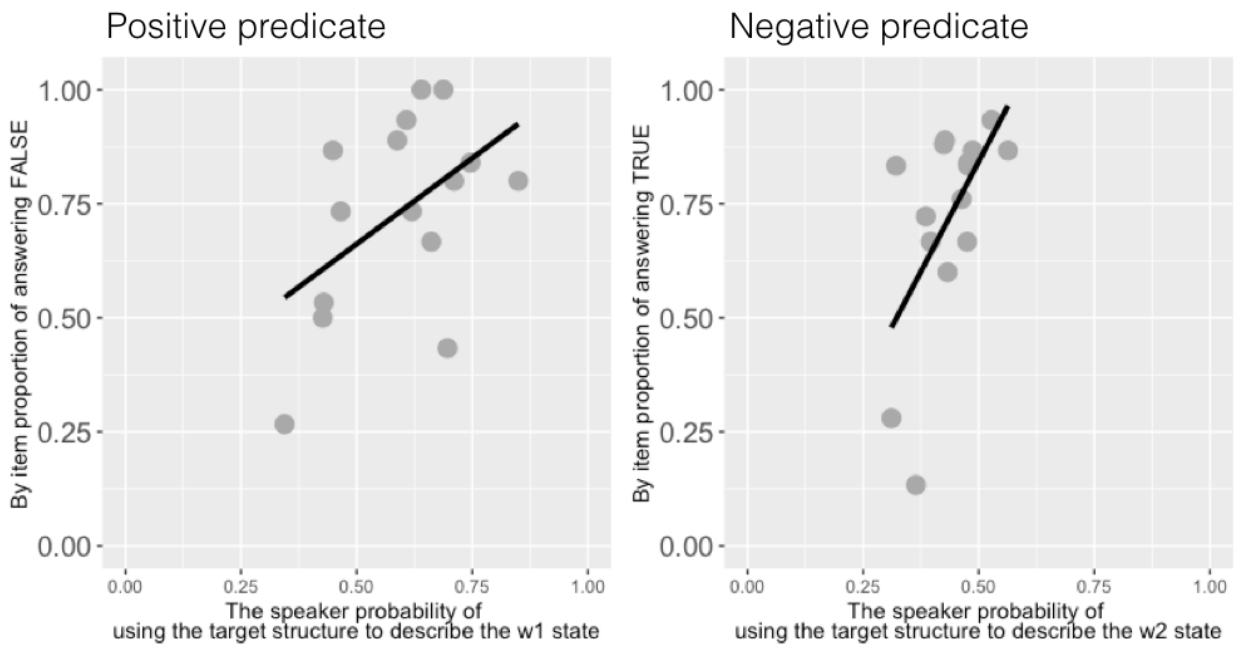


FIGURE 7. By-item correlation between truth value judgments and production bias. Left: positive predicates, $p = 0.059$; Right: negative predicates, $p = 0.02$.