Parsing and interpretation: When parsing decisions misalign with interpretive decisions

Ming Xiang *[1], Zhewei Dai[2], and Suiping Wang*[3]

[1]Language Processing Lab, Linguistics Department, University of Chicago, IL, 60615

[2]Department of Mathematics and Computer Science, Alma College, MI 48801

[3]Department of Psychology, South China Normal University, Guangzhou, China

*Corresponding Authors. Emails: mxiang@uchicago.edu, wangsuiping@scnu.edu.cn

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 representation preferred 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 is concerned with the locality bias in non-local dependencies, specifically, the Mandarin wh-in-situ scope dependencies. Our findings suggest a misalignment between the local parsing decisions and the global interpretative decisions. 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. Following the Rational Speech Act framework (RSA, Goodman & Frank, 2016), we propose a Bayesian pragmatic inference model to account for these findings. Under this model, the seeming conflict between parsing and interpretation will ultimately disappear because in the proposed model parsing preferences will be naturally embedded under the pragmatic reasoning process to derive the ultimate interpretation. The currently study therefore makes novel contributions, both empirically and theoretically, to address questions about the relationship between parsing and interpretation.

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 linguistic representation. It is often explicitly or implicitly assumed that the representation preferred by the parser would be mapped onto the ultimate interpretation of the sentence. In other words, the structure endorsed by the parser should determine the interpretation of the sentence. This traditional view is not unchallenged. For example, studies on *garden-path* sentences have revealed that comprehenders can obtain interpretations that conflict with the possible parse of the linguistic input (Christianson et al. 2001; Qian et al. 2018). The *good enough* approach to comprehension (Ferreira et al., 2001, 2002; Christianson et al. 2001; Ferreira & Patson, 2007) explains such findings by allowing interpretations derived through simple heuristics (e.g. world knowledge, word order, etc) to trump interpretations based on a fully specified parse. 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 *garden-path* phenomenon, that demonstrates (descriptively speaking) misalignment between parsing and interpretation. Second, our account of the misalignment offers a new kind of analytical possibility to address the general question about the relationship between parsing and interpretation. Specifically, we will argue that interpretation should be modeled as a pragmatic inference of the comprehender. We follow the Rational Speech Act framework (RSA, Goodman & Frank, 2016) and propose a Bayesian pragmatic inference model to account for our findings. The seeming conflict between parsing and interpretation will ultimately disappear because in the proposed model parsing preferences will be naturally embedded under the pragmatic reasoning process to derive the ultimate interpretation.

Our case study concerns with the locality bias in sentence processing. In particular, we will examine the locality effect in parsing and understanding scope ambiguity for wh-in-situ constructions. 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 longer 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 and Gibson 2002; Van Dyke and Lewis 2003; Lewis and Vasishth 2005). As an example, consider (1) from Alexopoulou and Keller (2007). In their results, with the distance between the verb *fire* and its fronted wh-argument *who* increasing 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 than 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 references between the two elements on a dependency chain. Under the memory retrieval account (Lewis and Vasishth 2005), the cost is significantly determined by how quickly and unambiguously the relevant dependent element can be retrieved from working memory, against all other memory representations that could potentially introduce interference. All the major accounts of locality bias are hypotheses about how structural representations are established. The term *structural* here should be interpreted in a broad sense, so it encompasses not only syntactic but also semantic structures as well, assuming that compositional semantics and syntactic structures closely track each other. Under these accounts, the locality effects are the natural consequence of a parsing architecture that employs a particular set of parsing procedures to incrementally incorporate each incoming word into the current syntactic and semantic representations.

Although there is ample discussion in the literature about the parsing mechanisms for completing non-local dependencies, there is relatively little discussion about how the parsing outcome maps to the interpretation a comprehender obtains. A simple hypothesis would be that if a comprehender adopts a particular parse, they would also adopt the interpretation this parse generates. We examine this question below using Mandarin wh-in-situ construction as our case study.

2. Parsing wh-in-situ scope – the locality bias

Some languages in the world are predominately wh-in-situ. Different from English, in wh-in-situ languages, wh-dependencies do not dislocate the wh-expressions to clause edge positions. Instead, wh-elements stay in their canonical theta positions. An example of a Mandarin Chinese wh-construction is given in (2):

(2)

记者们 知道 [$_{\text{Clause1}}$ 市长 严惩了 哪些官员。]

jizhemen zhidao shizhang yancheng-le naxie-guanyuan.

Reporter know mayor punish which-CL official.

“The reporters knew which officials the mayor punished.”

The example in (2) is interpreted as an embedded wh-question, as shown by its English translation. Even though the sentence has the word order like a declarative sentence, the wh-element *which official* ultimately takes scope on the left edge of the embedded clause, making the sentence an embedded question. In languages like English which regularly front their wh-elements in wh-constructions, the scope of a wh-element is explicitly signaled in the word order via the position of the wh-element, as demonstrated by the English translation above. In Chinese wh-in-situ dependencies, the relationship between a wh-element and its scope position could be analyzed in linguistic terms either as an abstract covert syntactic dependency or as a semantic dependency established in the semantic composition ((Aoun & Li, 1993; Cheng, 1991, 2003; Huang, 1982; Tsai, 1994).

Based on experimental evidence from eyetracking reading and acceptability ratings, Xiang et al. (2015; manuscript in progress) argued that when there is scope ambiguity for a wh-in-situ element, the more local scope dependency is less costly than the high scope dependency, essentially illustrating a locality bias like their wh-dependency kin in English. A slightly modified example from Xiang and colleagues is presented in (3):

(3) a.

记者们 知道 [$_{\text{Clause1}}$ 市长 透露了 [$_{\text{Clause2}}$ 市政府 严惩了 哪些官员。]]

jizhemen zhidao shizhang toulu-le shizhengfu yancheng-le naxie-guanyuan.

Reporter **know** mayor **reveal-perf** city-council punish-perf 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 the city council punished.” (low scope)

(3) b.

记者们 知道 [$_{\text{Clause1}}$ 市长 相信 [$_{\text{Clause2}}$ 市政府 严惩了 哪些官员。]]

jizhemen zhidao shizhang xiangxin shizhengfu yancheng-le naxie-guanyuan.

Reporter **know** mayor **believe** city-council punish-perf which-CL official.

“The reporters knew which officials the mayor believed that the city council punished.” (high scope)

NOT

“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 scope either at the left edge of clause 1 or the left edge of clause 2, as shown by the English translations, in which the scope of the wh-phrase is explicitly marked via the linear position of the wh-phrase in the sentence. The lower scope, i.e. the local scope dependency that associates the wh-item with the clause 2 boundary, was argued by Xiang and colleagues to be more preferred than 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. To see the critical verb differences between (3a) and (3b), let's consider examples in (4). Verbs like *know* or *reveal* allow either embedded interrogative or declarative clauses as their complements, 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). We will call the latter class of verbs *obligatorily -Q* verbs, with *Q* as an abbreviation for *questions.* We only used English examples in (4) to demonstrate the verb differences, but similar verb properties hold for Mandarin as well.

(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 lower scope dependency in (3b) is blocked, since constructing the lower scope dependency would result in forming an interrogative embedded clause for a verb like *believe.*

In Xiang and colleagues' results, blocking the local scope dependency in (3b) led to substantial processing difficulty, which resulted in much lower acceptability rating for (3b) than (3a) and longer regression reading time on the wh-morpheme in (3b) than (3a). The contrast between (3a) and (3b) strongly suggests a locality bias in identifying the scope position for wh-in-situ expressions. If interpretation bias is parallel to the parsing bias, we would expect that the scope ambiguity should ultimately be resolved to favor interpretations supported by the local dependency. We test this in Experiment 1 using a truth value judgment task.

3. Experiment 1: Scope interpretation bias – Truth value judgment task

In this experiment, participants were presented with a sentence containing a wh-in-situ expression. The target sentence by itself has two possible interpretations: one is compatible with the low scope (local) dependency, and the other with the high scope dependency. But the participants were also presented with a context scenario that was only compatible with one of the interpretations. They were instructed to judge whether the target sentence *fit* the context. Their judgments, therefore, can provide us with some hint as to which scope dependency they have committed to.

Consider a target sentence like the following:

(5)

艾米丽	公布了 [$_{\text{Clause1}}$	她的团队	发现了 [$_{\text{Clause2}}$	外星人	建造了	哪座城市。]]
Emily	gongbu-le	tade tuandui	faxian-le	waixingren	jianzhao-le	naxzuo-chengshi.
Emily	**announce-perf**	her team	**discover-perf**	aliens	establish-perf	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, it 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 difference between the high and low scope dependencies will play an important role in our experiment below.

3.1 Material, participants and procedure

We constructed four different conditions. The first condition (6a) has a target sentence like (5), and a preceding context that is incompatible with the high scope reading and compatible with the low scope reading. Participants were instructed to judge, after reading the context and the target sentence, whether the target sentence *fit* or *did not fit* the context scenario. For convenience, we will refer to the task as a truth value judgment task, and code the *fit* and the *did not fit* responses as *true* and *false* judgments respectively. As mentioned above, the context in (6a) makes the high-scope construal *true* and the low-scope construal *false* for the target sentence. To counterbalance the association between the True/False judgments and the high/low scope construal of the target sentence, we modified the matrix predicate in (6a) to create the condition (6b). In the condition (6b), the matrix verb is the antonym of the positive matrix predicate in (6a). In the majority of the stimuli items (12 items out of 16), the verb in (6b) contains an overt negation marker followed by a positive predicate. For convenience, we labeled the condition (6b) as "Matrix verb negative", even though not all of the items, including the example in (6b), contained a morphologically negative matrix predicate. The context for (6b) is identical to (6a), but because the matrix verbs in these two conditions were antonyms, we expect the judgments provided to the target sentence should be switched. In this way we counterbalanced the the association between the True/False judgments and the high/low scope construal of the target sentence.

(6a): Ambiguous; Matrix verb positive

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 didn't release the name of the city.

Target sentence:

艾米丽 公布了 [$_{\text{Clause1}}$ 她的团队 发现了 [$_{\text{Clause2}}$ 外星人 建造了 哪座城市。]]

Emily gongbu-le tade tuandui faxian-le waixingren jianzhao-le naxzuo-chengshi.

Emily **announce-perf** her team **discover-perf** aliens 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**)

(6b): Ambiguous; Matrix verb negative

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 didn't release the name of the city.

Target sentence:

艾米丽 隐瞒了 [$_{\text{Clause1}}$ 她的团队 发现了 [$_{\text{Clause2}}$ 外星人 建造了 哪座城市。]]

Emily yinman-le tade tuandui faxian-le waixingren jianzhao-le naxzuo-chengshi.

Emily **hide-perf** her team **discover-perf** aliens establish-perf which-CL city.

High scope: “Emily hid which city her team discovered aliens established.” (**True**)

Low scope: “Emily hid her team discovered which city the aliens established.” (**False**)

In addition to the two ambiguous conditions in (6a) and (6b), we also included unambiguous target sentences as control comparison conditions, see (6c) and (6d). For these conditions, the context scenarios are identical to the ambiguous target sentence conditions above, but the target sentences themselves were made unambiguously high scope.

(6c): Unambiguous; Matrix verb positive

Context: (the same as above)

Target sentence:

艾米丽	公布了 [$_{\text{Clause1}}$	她的团队	相信 [$_{\text{Clause2}}$	外星人	建造了	哪座城市。]]
Emily	gongbu-le	tade tuandui	xiangxin	waixingren	jianzhao-le	naxzuo-chengshi.
Emily	**announce-perf**	her team	**believe**	aliens	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." (**not available**)

(6d): Unambiguous; Matrix verb negative

Context: (the same as above)

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

High scope: "Emily hid which city her team believed aliens established." (**True**)

Low scope: #"Emily hid her team believed which city the aliens established." (**not available**)

The target sentences in (6c) and (6d) are unambiguous because the lower embedded verb *believe* is lexically unable to take an interrogative complement clause. The in-situ wh-phrase could not take scope at the edge of the lower clause, since "Emily's team believed which city the aliens established" is not a grammatically well-formed structure. As a result, only the high scope reading is grammatically available for the target sentences in (6c) and (6d). Given the context scenario preceding these two target sentences, (6c) should be judged as false, and (6d) true.

We constructed a total of 16 sets of 4-condition items like (6a-d). The experiment was conducted on the platform Ibex Farm. 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. They were instructed to decide whether the target sentence *fit* or *does not fit* the given context, and make their choice by clicking one of the 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, and the other 5 as false. Ninety-eight native Mandarin speakers participated in our study, 10 of which were excluded because their response accuracy on the filler trials was lower than 60%. We report the results from the remaining 88 participants below.

3.2 Results

We first converted participants's truth value judgments into whether they interpreted the target sentence with a high scope construal. For example, for (6a), a response of *false* was converted to *high scope*. The proportion of high scope construal is plotted in Fig. 1. There are slightly more high scope construals for the unambiguous conditions than the ambiguous conditions ($Est = -0.21 \pm 0.08, z = -2.58, p < .01$). This is not surprising given that the unambiguous conditions can only be parsed as having a high scope for the wh-expressions. What is more interesting is that the two ambiguous conditions both received overwhelmingly more high-scope construal, 79% for the positive predicate condition and 77% for the negative predicate condition. Both are significantly higher than the 50% chance level (*ps* $<$ .0001).

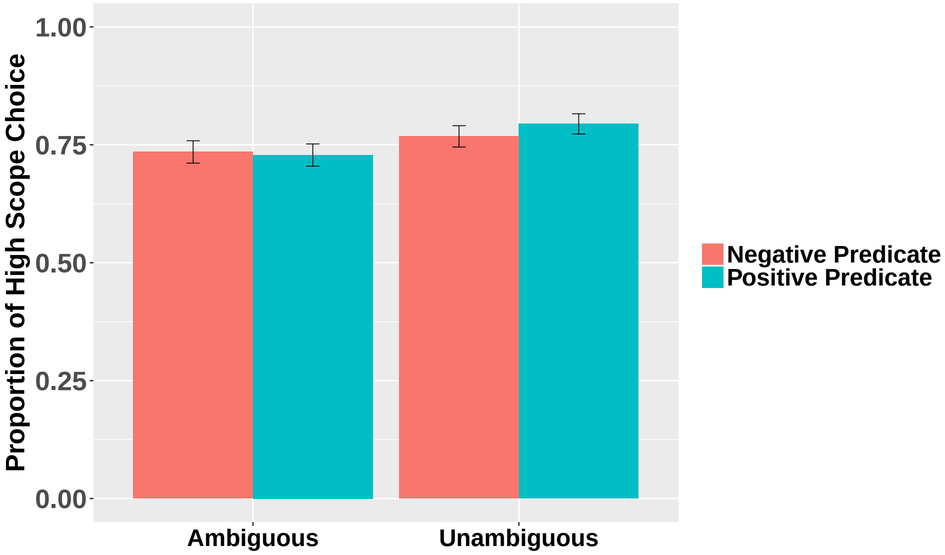

Fig 1: Truth value judgment task results: proportion of participants' choosing the high scope construal

3.3 Discussion of Experiment 1

Results from the truth value judgment task in Experiment 1 provided strong evidence that participants are predominately biased towards interpreting an ambiguous wh-in-situ construction as having a high scope reading. This finding contradicts any prediction based on a positive correlation between interpretation bias and parsing complexity. 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 long distance 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.

The conclusion that the interpretation bias tested here is the opposite of the parsing bias criticially depends on previous findings in Xiang et al. (2015b, 2018). 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. In particular, a context scenario was included in the current study, but was absent in the previous study. The verbs used in these two studies are also not entirely identical. Experiment 2 aims to replicate the parsing locality bias using the current set of stimuli. Since both acceptability rating and eye-tracking results converged on the locality bias in Xiang and colleagues' findings, we use an acceptability judgment task in Experiment 2 to assess the parsing bias.

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

4.1 Material, participants, procedure and predictions

The material for Experiment 2 was identical to Experiment 1, with a total of 16 sets of 4-condition experimental items (an example in 6a-d) and 10 filler items. The experimental procedure was also identical to Experiment 1, except that 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 to favor the local scope dependency, we expect to replicate the acceptability findings in Xiang et al. (2015b, 2018). For ambiguous conditions (6a) and (6b), the local dependency is available, and these sentences should be judged as acceptable once the local dependency is successfully constructed. For the unambiguous conditions (6c) and (6d), on the other hand, the local dependency is blocked. We predict lower acceptability ratings on these unambiguous sentences since constructing a non-local high scope dependency is costly for the parser.

4.2 Results

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 than the ambiguous conditions regardless of whether the predicate was positive or negative ($Est = -1.01 \pm 0.37, z = -2.78, p < .01$). Sentences with positive matrix predicates were also rated lower than those with negative matrix predicates ($Est = -0.82 \pm 0.34, z = -2.44, p < 05$). We did not have any hypothesis/predictions for the predicate difference, so we will not discuss the effect of predicate any further.

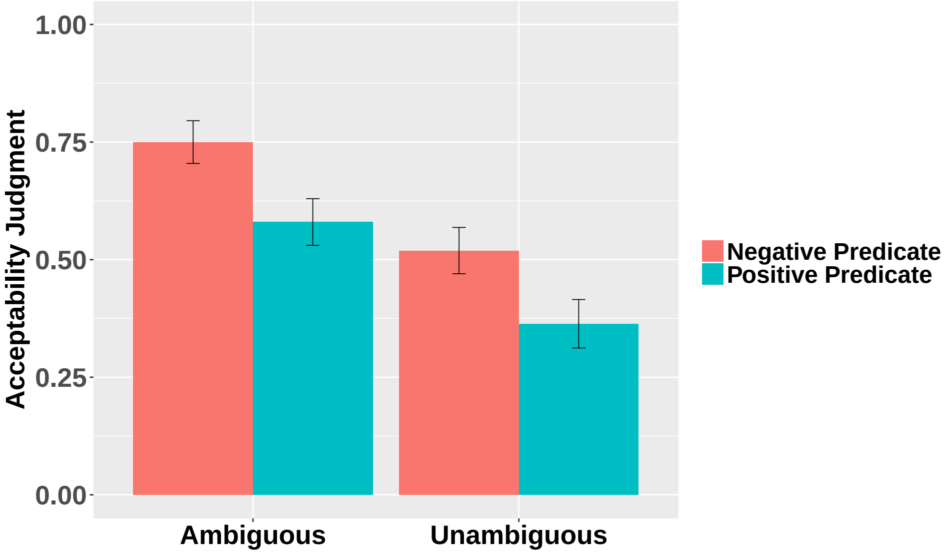

Fig. 2: Acceptability judgment results: Proportion of participants' Yes responses

4.3 Summary and discussion

The results from Experiment 1 and 2 present an empirical paradox. On the one hand, consistent with previous eye-tracking and acceptability rating results, Experiment 2 confirmed the locality bias in parsing a wh-in-situ dependency. Sentences that made a local scope dependency available were judged much more acceptable than sentences that blocked the local scope dependency and allowed only a non-local high scope dependency. Importantly, additional follow-up experiments in Xiang et al. (2015b, 2018) also showed that the low acceptability for high-scope only sentences such as (6c) and (6d) was indeed due to the unavailability of the low scope dependency, instead of the potential alternative account that (6a) and (6b) may have simply benefited from the fact that they are ambiguous (e.g. the ambiguity advantage effect, Traxler, Pickering & Clifton, 1998). Experiment 1, however, showed an anti-locality bias in the ultimate interpretation participants obtained from scope ambiguous wh-in-situ constructions. These two experiments used exactly the same set of material, and the only difference was the task. It is well-known that processing complexity affects acceptability ratings: sentences that are difficult to parse often lead to reduced acceptability judgments (Chomsky & Miller 1963; Hofmeister et al., 2013). The acceptability results from Experiment 2 are therefore in line with previous findings, under the assumption that it is easier to parse a local scope dependency than a non-local one. The truth value judgment task in Experiment 1, on the other hand, is designed to probe the actual interpretation participants obtain from the linguistic input. The apparent contrast between the results from these two experiments raises the important issue that parsing biases do not necessarily align with interpretive biases.

Language comprehension requires constructing structural representations for the linguistic input. This is the task of parsing. Every piece of incoming linguistic input needs to be parsed into a structural representation. In the current case, a complete parse needs to specify where the scope position is for the wh-in-situ phrase. Needless to say there is a close relationship between the parsing and interpretation, since semantic composition needs to be based on the parsed structures. But in the mean time, comprehension is also affected by other parsing-independent processes, among which pragmatic reasoning has been recognized as one of the most important factors. The general idea that language comprehension should be viewed as the output of a cooperative process between speakers and listeners, involving sophisticated pragmatic reasoning, is an old and extremely influential one (Grice, 1975). In recent years, this idea has been formalized using Bayesian computational models (Goodman and Frank, 2016). 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 should pragmatically reason about the most likely messages the speaker has intended, given the form of the utterance, the context, and the listener's own world knowledge. We will first introduce some general background on the Bayesian pragmatic model. Given the model architecture, in order to evaluate whether the model prediction matches our truth value judgment results in Experiment 1, we will also present two additional experiments, one on participants' world knowledge, and the other on their production bias.

5. Modeling the truth value judgment results with Bayesian pragmatic inferences

5.1 Pragmatic inferences in language communication

Linguistic utterances convey information about the world. A pragmatic listener, upon hearing an utterance, would update their probabilistic model of the world states based on the information conveyed by the utterance. Following the recent rational speech act framework (Goodman and Frank, 2016; Frank and Goodman, 2012), we can model a pragmatic listener's belief about the world state w given the utterance u, using the Bayesian inference, as shown in equation (7):

$$(7)\ \ P_L(w|u) = \frac{P_S(u|w) \times P_L(w)}{\sum_{w'} P_S(u|w') \times P_L(w')}$$

The pragmatic listener (L) is conditioning their belief update on two factors. First, assuming the speaker S is cooperative and trying to be helpful, the listener works backwards and estimate the likelihood a speaker would have uttered u given the world state in the speaker's mind. Second, the listener also brings to the communication some prior belief as to how likely the world state w holds independent of the utterance. The normalizing constant in (7) (i.e. the denominator), considers the alternative world states that the listener entertains as relevant when decoding the message delivered by the utterance u. In this kind of framework, therefore, in order to model a pragmatic listener, we also need to model a pragmatic speaker, who chooses to utter u among a set of alternative utterances to describe a world state in her mind. The RSA model, therefore, captures the intuition that linguistic communication involves back-and-forth pragmatic reasoning between a listener and a speaker.

For the current purpose, our main interest are the ambiguous utterances in (6a) and (6b), and the variable u corresponds to these utterances. These examples are repeated below in (8a) and (8b), but only with English glosses in the Chinese word order. Since these utterances are ambiguous, and could convey information about different world states, the listener's task is to infer the probability of each possible w given u.

(8a) Positive matrix predicate

Emily *announced* her team *discovered* aliens established which city.

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

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

(8b) Negative matrix predicate

Emily *hid* her team *discovered* aliens established which city.

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

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

It is important to be clear what the relevant world states w could be for the listener. 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 (but not necessarily) correspond to more than one state in the world (see more discussion about this in section 5.3). Let's first make clear what the relevant world states are for our working example in (8). When the matrix predicate is positive, as in (8a), the relevant world states are a set of possible combinations of two events: e1: Emily announced the name of a city, which they discovered was built by aliens; and e2: Emily announced their discovery that they found out which city 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 utterance shouldn't have been made by the speaker 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

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

in (8b), the relevant worlds states are also a set of possible combinations of two events: the *name hiding* e1: Emily hid the name of a city, which they discovered was built by aliens; and the *discovery hiding* e2: Emily hid their discovery that they found out which city 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 hide the discovery of the city without also hiding 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.

Table 2: World states relevant for utterances with negative predicate

world states	e1 name hiding	e2 discovery hiding	Considered as a relevant world state?
w_1	$+$	$+$	yes
w_2	$+$	$-$	yes
w_3	$-$	$+$	no
w_4	$-$	$-$	no

In Table 3, we summarize the remaining possible 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 the same; but their corresponding w_1 states are different.

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 which city was built by aliens and she also announced the name of the city.	Emily hid the fact that they discovered which city was built by aliens and (necessarily) also hid the name of the city.
w_2	Emily announced they discovered which city was built by aliens but she did not announce the name of the city.	Emily did not hide the fact that they discovered which city was built by aliens but she hid the name of the city.

Based on the equation in (7), in order to compute $P_L(w_1|u)$ or $P_L(w_2|u)$, one also needs to know the prior probability for each world state, and the likelihood for the speaker to produce the target utterance given the world state they have in mind. We will empirically estimate the prior probability in Experiment 3 below. For the speaker likelihood, we will make two different estimates. First, the RSA framework provides the general principles for us to make a model prediction about the speaker behavior. In addition, we will also empirically estimate speakers' production bias in Experiment 4.

In the RSA framework, the pragmatic speaker is 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). The pragmatic speaker would in general want to maximize her utility. The utility function could be defined in a number of ways (Goodman and Frank, 2016), and we follow the most basic definition that captures the intuition that a helpful speaker would choose to make the most informative utterance to the listener, as shown in (10). The equation in (10) measures how certain a listener is for a particular world state w upon hearing u. Intuitively, if the speaker makes a very informative utterance, the listener should update her beliefs in such a way that the world state w intended by the speaker would become more likely in the listener's posterior beliefs. To avoid infinite recursion, the listener L_0 in (10) is defined to be a simple *literal* listener, who updates their beliefs based on whether the literal meaning (i.e. the semantic meaning) of the utterance is true, as shown in (11).

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

(10) $U_S(u; w) = ln(L_0(w|u))$

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

The literal listener L_0 in (11) is crucial in order to connect the pragmatic reasoning process to the compositional semantics of the linguistic input. The truth value (1 or 0) of an utterance is determined by considering whether the utterance $[\![u]\!]$ is true or false when applied to 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 by renormalizing the remaining world states (i.e. the ones that are compatible with the semantics of the utterance) based on their prior probabilities.

We will demonstrate in more detail below how each of these steps can be implemented for the current empirical case. To start, since prior probabilities for the relevant world states play an important role in the model prediction, we conducted an experiment to empirically estimate the prior probabilities for the different world states in Table 3.

5.2 Experiment 3: Estimating the prior probabilities

5.2.1 Material, participants and procedure

To experimentally assess the prior probabilities of each different world state relevant for the listener, we first provided participants a neutral context that corresponds to the background scene used in Experiment 1, for example, a background scene about an archeology conference. Participants were then instructed to choose between two possible situations that could take place in the given context. These two situations corresponded to the two different world states illustrated in Table 3, although 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 based on material from Experiment 1. Sixteen sets of items were constructed corresponding to the original 16 sets of scenarios in Experiment 1, each with 2 conditions. An example is given in (12) below:

(12) Context: At a recent archaeology conference, Emily made a presentation on behalf of her research team.

Question: Which of the following situation is likely to arise?

(12a) For the positive predicates:
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.
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 as a secret for the moment.

(12b) For the negative predicates:
w_1: Emily's research team actually have 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.
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 as a secret for the moment.

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 items.

5.2.2 Results

Among the choices participants made, for the *positive predicates* condition, there was a slight preference for the w_1 state (0.53 w_1 vs. 0.47 w_2), marginally different from the chance performance ($p = 0.07$); 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 ($p < 0001$). With the prior probabilities estimated, we are ready to model the literal listener L_0, the pragmatic speaker S and ultimately the pragmatic listener L.

5.3 The literal listener model and the pragmatic speaker model

As mentioned earlier, the literal listener L_0 is the crucial step in the Bayesian pragmatic model that connects structured semantic composition to pragmatic reasoning. In the current case, the compositional semantics of the utterance u depends on how the surface string is parsed into different structures with different scope dependencies. Let's call the two dependency parses u_h and u_l, standing for high-scope parse and low-scope parse. Because of the ambiguity in parsing possibilities, upon hearing the utterance, L_0's inferences need to be weighted by the probability of each parse, as shown in (13), which also incorporates the equation from (11):

(13) $L_0(w)$

$$= L_0(w|u_h) \times P(u_h) + L_0(w|u_l) \times P(u_l)$$
$$= \frac{\delta_{[\![u_h]\!](w)} P(w)}{\sum_{w'} \delta_{[\![u_h]\!](w')} P(w')} \times P(u_h) + \frac{\delta_{[\![u_l]\!](w)} P(w)}{\sum_{w'} \delta_{[\![u_l]\!](w')} P(w')} \times P(u_l)$$

Let's consider our working example in (8a), repeated in (14), in which the matrix predicate is positive. For convenience, We also repeat from Table 3 the two relevant world states assumed for this utterance.

(14) Emily *announced* her team *discovered* aliens established which city.

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

w_1 *positive*: Emily announced they discovered which city was built by aliens and she also announced the name of the city.
w_2 *positive*: Emily announced they discovered which city was built by aliens but she did not announce the name of the city.

To compute (13), we will first make the simple assumption that it is equally likely to parse the ambiguous string in (14) 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 in the end. If the utterance u is parsed as u_h, it specifies the fact that the name of the city was made known, and therefore we obtain a truth value 1 with w_1, but 0 (i.e. false) with w_2. If the utterance u is parsed as u_l, since it underspecifies whether the name of the city is made known, we could not remove either w_1 or w_2 from consideration, therefore we keep both 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 (Experiment 3). The literal listener L_0 therefore updates her beliefs about w_1 and w_2 in the following way:

(15) $L_0(w_1|u_{positive})$

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

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

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

$$= 0.765$$

$L_0(w_2|u_{positive})$

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

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

$$= 0 + 0.47 \times 0.5$$

$$= 0.235$$

This computation suggests that even though the 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 (14), the listener is leaning much more towards believing in w_1 over w_2.

When the utterance contains a negative matrix predicate (our working example is repeated in 16), the computation is still based on (13), but the truth conditions will change. When the utterance u is parsed as u_h, it is compatible with both w_1 and w_2, and hence both states need to be considered by the listener. If the utterance u 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.

(16) Emily *hid* her team *discovered* aliens established which city.

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

w_1 *negative*: Emily hid the fact that they discovered which city was built by aliens and (of course) also hid the name of the city.
w_2 *negative*: Emily did not hide the fact that they discovered which city was built by aliens but she hid the name of the city.

(17) $L_0(w_1|u_{negative})$

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

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

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

$$= 0.71$$

$L_0(w_2|u_{negative})$

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

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

$$= 0.58 \times 0.5 + 0$$

$$= 0.29$$

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

With the estimates for the L_0 listener, we can estimate how likely a pragmatic speaker would utter u to describe a particular world state w in her mind. Based on equations in (9) and (10), the pragmatic speaker is predicted to have the following behavior:

(18) $P_S(u|w) \propto exp(\alpha \times ln(L_0(w|u))$

The free parameter α in (18) captures the extent to which the speaker is a rational agent, i.e. how much she would choose her utterance to maximize her utilities, which in (18) is the informativity of the utterance to the literal listener. We won't set a particular value for α here. But it should be clear from (18) that if the posterior belief of L_0 favors w_1, as shown by the calculations above for both positive and negative predicates, the pragmatic speaker is more likely to choose the ambiguous target string in (14) and (16) to describe w_1 than w_2.

The calculations above have been based on the assumption that the literal listener has no parsing bias to parse an ambiguous string u into either a high-scope or a low-scope dependency. This assumption needs refinement, since we already know there should be parsing bias favoring the low scope dependency. After we introduce the constraint $0 < p(u_h) < 0.5$ and $0.5 < p(u_l) < 1$ into the calculation in (15) and (17), it could be derived that for utterances with positive predicates (15), the literal listener's inferred probability for w_1 is between 0.53 and 0.765; and for utterances with negative predicates (17), it is between 0.71 and 1. In other words, given the parsing bias that favors the low-scope dependency, the literal listener is predicted to always favor w_1 over w_2. Consequently, the model will predict that the pragmatic speaker is always more likely, for both types of predicates, to produce the ambiguous utterance when describing w_1 than describing w_2. In Experiment 4 below, we empirically validate whether the model prediction for the pragmatic speaker bears out.

5.4 Experiment 4: Empirically estimating the production bias

5.4.1 Material, procedure and participants

The goal of this experiment is to estimate how likely participants will use the target wh-in-situ construction to describe a particular world state. To this end, we first constructed scenarios that correspond to the world states relevant for our purpose. We adapted the paraphrases of different world states from Experiment 3, as shown by the examples in (12a) and (12b). As noted earlier, these world states correspond to the ones described in Table 3, but with different paraphrases. 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 (6a) and (6b). 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, Wang and Cui (2015a) showed that native Mandarin speakers avoided producing wh-dependencies as much as they can, even at the cost of producing some otherwise dis-preferred 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.

The experimental trials have the following structure. First, participants were given one of the four world state scenarios from (12a) and (12b). Then they were given four phrase fragments. They were instructed to form a sentence, using these fragments and any additional material they want to use, that expresses a message coherent with the context scenario presented to them. The four fragments were presented in a 2x2 grid format, and the position of each fragment in this grid was randomized from trial to trial. For example, if a participant received a world state scenario for the positive predicate (i.e. one of the two world states under (12a)), the four fragments they would receive were *"Emily announced"*, *"which city"*, *"established"*, *"her team discovered"*. If a participant received a relevant world state scenario for the negative predicate (i.e. one of the two world states under (12b)), they would receive the almost identical set of fragments except that the positive predicate will be changed to the negative one *"Emily hid"*. The positions of these fragments in the 2x2 grid were randomized so that participants won't be cued about the word order of the target sentence they will produce. Even though the task itself is not equivalent to spontaneous natural production, we think it nevertheless leaves participants enough flexibility to form various types of utterances and they were not overly 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 (12), 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.

5.4.2 Results

Three different native Mandarin speakers coded the production results. We removed the trials (about 1% of the total trials) from participants that didn't perform the task properly. 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.

Among all trials, 40% of them conformed to the wh-in-situ target structure, with the same word order as the target sentences used in Experiment 1. In fig. 3 we present the proportion of producing the target structure split by world state context and the predicate type. For both types of predicates, participants were more likely to produce the ambiguous target structure when describing world state 1 than when describing world state 2 ($Est = 0.32 \pm 0.05, z = 6.93, p < .00001$).

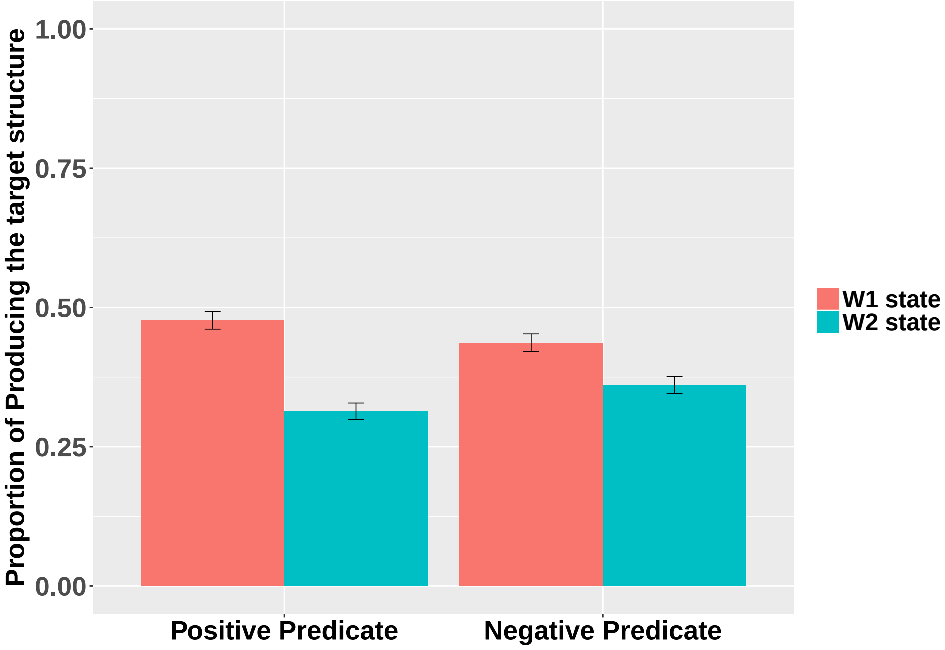

Fig 3: Proportion of producing the target wh-in-situ structure

5.4.3 Summary of Experiment 4

The empirical production results confirmed the model prediction in the last section: for both positive and negative predicates, speakers preferred to use the target ambiguous wh-in-situ structure when they were describing scenario w_1; and they produced the target structures less frequently when they were describing scenarios corresponding to w_2.

5.5 Deriving the pragmatic listener's inferences

Finally, we are ready to work out the model prediction for the posterior beliefs for the pragmatic listener. When a pragmatic listener hears an ambiguous utterance containing a wh-in-situ expression, how would she update her beliefs about the world? Equation (7) is repeated in (19) below.

$$(19)\ \ P_L(w|u) = \frac{P_S(u|w) \times P_L(w)}{\sum_{w'} P_S(u|w') \times P_L(w')}$$

To compute (19), we need to plug in the empirical production results from Experiment 4 and the prior probability results from Experiment 3. For convenience, we first summarized the results from these two experiments in Table 4 and 5, for the positive and negative predicates separately, and then demonstrated how listener posterior probability is calculated.

Table 4: For the positive predicate, see the example in (6a) and (8a):

Emily *announced* her team *discovered* aliens established which city.

world states	$P_s(u|w)$	Priors
w_1: Emily announced they discovered which city was built by aliens and she also announced the name of the city.	0.48	0.53
w_2: Emily announced they discovered which city was built by aliens but she did not announce the name of the city.	0.31	0.47

(20)

$$\begin{aligned}
P_{L_1}(w_1|u_{positive}) &= \frac{P_{S1}(u|w_1) \times P_{L_1}(w_1)}{\sum_{w'} P_{S1}(u|w') \times P_{L_1}(w')} \\
&= \frac{0.48 \times 0.53}{0.48 \times 0.53 + 0.31 \times 0.47} \\
&= 0.64 \\
P_{L_1}(w_2|u_{positive}) &= \frac{P_{S1}(u|w_2) \times P_{L_1}(w_2)}{\sum_{w'} P_{S1}(u|w') \times P_{L_1}(w')} \\
&= \frac{0.31 \times 0.47}{0.48 \times 0.53 + 0.31 \times 0.47} \\
&= 0.36
\end{aligned}$$

The pragmatic listener therefore believes w_1 is more likely upon hearing the target utterance. Under the truth value judgment task in Experiment 1, see example (6a), if a participant's posterior belief upon hearing the utterance favors w_1, she will answer *False*, because the w_1 state is contradicting what the context scenario describes in (6a). A response of *False* to (6a) is only compatible with the high-scope parse, but not the low-scope parse of the target sentence in (6a), correctly deriving why the participants had a preference for the high-scope compatible interpretation of (6a).

We next performed the same calculation for utterances containing a negative predicate, such as (6b).

Table 5: For the negative predicate, see the example in (6b) and (8b).

Emily *hid* her team *discovered* aliens established which city.

world states	$P_s(u\|w)$	Priors
w_1: Emily hid the fact that they discovered which city was built by aliens and (necessarily) also hid the name of the city.	0.44	0.42
w_2: Emily did not hide the fact that they discovered which city was built by aliens but she hid the name of the city.	0.36	0.58

(21)

$$\begin{aligned}
P_L(w_1|u_{negative}) &= \frac{P_S(u|w_1) \times P_L(w_1)}{\sum_{w'} P_S(u|w') \times P_L(w')} \\
&= \frac{0.44 \times 0.42}{0.44 \times 0.42 + 0.36 \times 0.58} \\
&= 0.47 \\
P_L(w_2|u_{negative}) &= \frac{P_S(u|w_2) \times P_L(w_2)}{\sum_{w'} P_S(u|w') \times P_L(w')} \\
&= \frac{0.36 \times 0.58}{0.44 \times 0.42 + 0.36 \times 0.58} \\
&= 0.53
\end{aligned}$$

The pragmatic listener therefore believes w_1 is less likely, and w_2 is more likely, upon hearing the target utterance containing a negative predicate. Using our working example (6b), if if a participant's posterior belief upon hearing the utterance favors w_2, she will answer *True*, because the w_2 state is consistent with the context in (6b). Since a response of *True* is only compatible with the high-scope parse but not the low-scope parse of the target sentence in (6b), we also correctly derive the participants' preference for the high-scope compatible interpretation of (6b).

The pragmatic listener's updated posterior probabilities of the world, as predicted by the Bayesian pragmatic model, are therefore correctly predicting the truth value judgment responses observed in Experiment 1. But we want to note that even though the qualitative patterns of Experiment 1 results are well-predicted by our calculation in (20) and (21), quantitatively speaking, the empirical effects from Experiment 1 are more pronounced than what is predicted: 73% empirical vs. 64% predicted *False* responses for the positive predicate condition in (6a), and 73% empirical vs. 53% predicted *True* responses for the negative predicate condition in (6b). Since the model prediction in (20) and (21) made use of empirically collected estimates for production bias (Experiment 4) and priors (Experiment 3), it is likely that there is a substantial amount of noise in our empirical estimates, which affected the accuracy of model prediction. But the relatively large difference between the predicted and the empirical estimates for those utterances containing negative predicates seems to invite explanations beyond simply noise in the data. We do not have a fully developed account for this, but we will discuss a possible account in the General Discussion section.

6. 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 that was not previously observed. When wh-scope ambiguity is concerned for a wh-in-situ language like 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, following the rational speech act framework (Frank and Goodman, 2012; Goodman and Frank, 2016), could provide a principled (at least partial) explanation of the interpretation bias, while also incorporating the parsing bias into the model. During the process of constructing the Bayesian pragmatic model, we also showed that the model could deliver correct qualitative predictions for the production of wh-in-situ constructions.

Our findings highlight the difference between the local parsing decisions and the global interpretative decisions. The goal of the parser is to establish a structural representation of the linguistic input. For each incoming word/phrase to be incrementally integrated into a structural representation, the parser may be affected by a number of factors, including the complexity of to-be-established structure (Frazier and Fodor, 1978), working memory constraints (Gibson, 1998; Lewis, Vasishth and Van Dyke, 2006), syntactic or semantic expectation of the upcoming material (Hale, 2003; Levy, 2008), or pragmatic principles (Tanenhaus et al., 1995). Once the linguistic input is parsed, in many cases the parsing output is parallel to the ultimate interpretation comprehenders obtain; but as shown by the current results, this is not necessarily the case. The gap between the local parsing biases and global interpretation biases is not surprising, however, if we consider a comprehender's task as not only structurally representing the incoming utterance, but more importantly inferring a message based on the communicative context the utterance was made. While recognizing the important role of pragmatic reasoning in comprehension, it is equally important to recognize how the Bayesian pragmatic model we adapted in this study incorporates the parsing preferences as part of processes that derive the comprehender's pragmatic inferences. In the architecture of the model, the pragmatic speaker's utility function is largely determined by the *literal listener*, which is the crucial step for structured representations and the compositional semantics computed over these structures to be considered by the comprehender. In the current case, as we showed in the calculation in (13), the truth conditions were calculated based on the parsed structure; and most importantly, the parsing bias was incorporated at this stage. In this sense, the interpretation bias in the end is not *contradicting* the parsing bias per se, since the latter was actually an inherent part of the computation contributing to the ultimate interpretation.

We see the incorporation of structured parses into the pragmatic reasoning process as an important feature of the RSA model we adapted here. It makes the RSA model an appealing candidate to potentially address a broader range of questions related to parsing and interpretation. As we mentioned in the introduction, findings from other empirical domains, such as the *garden-path* sentences, have already motivated different models that address the observed misalignment between parsing and interpretation. The empirical problem examined in this study is very different from the traditional garden-path sentences, but it is suggestive of a new analytical possibility for dealing with the general question. It is worth exploring in future research whether these different empirical problems can all be explained using one unified model.

Although the Bayesian pragmatic model provided good qualitative predictions for the interpretation bias, as we noted earlier (section 5.5), the quantitative fit was not the best. The mismatch was more salient when the utterance contained a negative matrix predicate – the model predicted the *true* response for examples like (6b) to be only slightly above chance (53%), whereas the empirical result was substantially above chance (73%). This discrepancy suggests to us the current model architecture needs further refinement. We speculate here that making the model more sensitive to what are the relevant questions under discussion (QUD, Ginzburg, 1996; Roberts, 1996) could be a possible model improvement in the future. 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 to the previous discourse context by virtue of helping to address the currently shared (often implicit) QUD at that given moment in time, either by providing an answer to it or by raising another relevant question, or via other discourse strategies. The comprehender would approach a given utterance as an answer to a discourse-salient QUD, and her pragmatic inference should be conditioned by this currently relevant issue. The difficulty with this approach is that QUDs are often assumed to be implicit, and there is no currently known rigorous method to identify them in a discourse context. This is also why we did not attempt to include them in our current model. Nonetheless, it seems possible that introducing QUDs into the model could potentially help refine the model predictions, especially for utterances containing negative predicates. Recall that in our working example (6), the context scenario ended with a note that Emily didn't announce the name of the city in their discovery. For the sake of argument, let's hypothesize that this last sentence was meant to answer an implicit QUD like "Did Emily announce the name of the city?". If this is the most recent/salient *issue/QUD* the context scenario was about, then when a comprehender received the target sentence and was asked to judge whether the target sentence fit with the context scenario, she may have focused only on this QUD and based her *true/false* judgments on how the target sentence answered this question. We have computed, in (20) and (21), the pragmatic listener's posterior beliefs about different world states after receiving an utterance. It is crucial to note that for an utterance containing a positive predicate, the two relevant world states in Table 4 would provide different answers to the QUD "Did Emily announce the name of the city?". The pragmatic listener was estimated to strongly bias towards w_1 in Table 4, which 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 the comprehender would be predicted to judge that the target sentence *does not fit* or *is false* under the given context. On the other hand, for an utterance containing a negative predicate, the two relevant world states in Table

5 would both trigger the same answer "*No, she didn't*" to the QUD, regardless of the listener's preferences for these two world states. This would mean that a comprehender should always conclude that the target sentence answered the QUD in a way consistent with how the QUD was resolved in the context, and therefore the sentence has a very high probability to be judged *fit* or *true* under the context.

Introducing an implicit QUD into our system, therefore, seems to provide a promising direction to better model the truth value judgment results. A number of previous studies have explored how to incorporate QUDs into the RSA models (Degen & Goodman, 2014; Savinelli, Scontras & Pearl, 2018; Scontras & Goodman, 2017), since it indeed seems to be a productive way to characterize the goal-driven discourse structure. But as we mentioned earlier, the major difficulty, at least for the current empirical domain, is to identify implicit QUDs in principled ways instead of stipulating what they are. We therefore will offer the suggestion above only as a speculation, and leave further model refinement for future work.

To conclude, our study provided novel empirical evidence to show that parsing and interpretation decisions can misalign with each other. We also showed that a model that approaches interpretation as Bayesian pragmatic inferences can incorporate parsing output into a unified 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, & Frank Keller. (2007). Locality, cyclicity, and resumption: At the interface between the grammar and the human sentence processor. *Language*, 2007, 110-160.

Aoun, J., & Li, A. Y.-H. (1993). Wh-elements in situ: Syntax or LF? *Linguistic Inquiry*, 24, 199-238.

Cheng, L. L-S. (1991). On the typology of wh-questions. PhD dissertation, MIT.

Chomsky, N., & Miller, G. (1963). Introduction to the formal analysis of natural languages. In R. Robert, R. Bush, D.Luce, & E. Galanter (Eds), *Handbook of mathematical psychology*, Vol.2, 269-321.

Christianson, K., Hollingworth, A., Halliwell, J. F., & Ferreira, F. (2001). Thematic roles assigned along the garden path linger. *Cognitive Psychology*, 42, 368–407.

Degen, J., & Goodman, N. D. (2014). Lost your marbles? The puzzle of dependent measures in experimental pragmatics. *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (pp. 397-402).

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

Ferreira, F., Bailey, K. G. D., & Ferraro, V. (2002). Good-enough representations in language comprehension. *Current Directions in Psychological Science*, 11, 11–15.

Ferreira, F., Christianson, K., & Hollingworth, A. (2001). Misinterpretations of garden-path sentences: Implications for models of sentence processing and reanalysis. *Journal of Psycholinguistic Research*, 30, 3–20.

Ferreira, F., & Patson, N. D. (2007). The “good enough” approach to language comprehension. *Language and Linguistics Compass*, 1, 71–83.

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

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

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

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

Gibson, E., Bergen, L., & Piantadosi, S. T. (2013). Rational integration of noisy evidence and prior semantic expectations in sentence interpretation. *Proceedings of the National Academy of Sciences*, 110, 8051–8056. doi: 10.1073/pnas.1216438110

Ginzburg, Jonathan. (1996). Dynamics and the semantics of dialogue. Seligman, Jerry, Westerst ahl, & Dag (Eds.), *Logic, language and computation*, 1.

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

Grice, H. Paul.(1975). *Logic and conversation.* In Peter Cole & Jerry Morgan (eds.), *Syntax and semantics*, vol. 3: Speech Acts, 43–58. New York: Academic Press.

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

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

Hofmeister, P., Jaeger T.F. , Arnon I., Sag IA, & Snider N. (2013). The source ambiguity problem: Distinguishing the effects of grammar and processing on acceptability judgments. *Language and Cognitive Processes* 28.1-2, 48-87.

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

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

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

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, 29, 375-419.

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

Qian, Z., Garnsey, S., & Christianson, K. (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, K. J., Scontras, G., & Pearl, L. (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* (pp. 67-75).

Scontras, G., & Goodman, N. D. (2017). Resolving uncertainty in plural predication. *Cognition*, volume 168, 294-311.

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

Traxler, M. J., Pickering, M. J., & Clifton, C. (1998). Adjunct attachment is not a form of lexical ambiguity resolution. *Journal of Memory and Language*, 39, 558-592.

Tsai, W.-T. D. (1994). On economizing the theory of A-bar dependencies. Ph.D. dissertation, MIT.

Van Dyke, J. A., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account ofrecovery from misanalyzed ambiguities. *Journal of Memory and Language*, 49, 285-316.

Warren, T., & Gibson, E. (2002). The influence of referential processing on sentence complexity. *Cognition*, 85, 79-112.

Xiang, M, Wang, S., & Cui, Y. (2015a).Constructing covert dependencies-The case of Mandarin wh-in-situ dependency. *Journal of Memory and Language*, 84, 139-166.

Xiang, M. Mo, L.& Wang, S. (2015b). Constructing wh-in-situ dependencies. *Presentation at Architectures and Mechanisms for Language Processing*, Valletta, Malta.

Xiang, M., & Wang, S.P (2018) Locality and Expectation in Chinese wh-dependencies. manuscript in preparation.