Processing Covert Dependencies: An SAT Study on Mandarin Wh-in-situ Questions

Ming Xiang[a]*, Brian Dillon[b], Matt Wagers[c], Fengqin Liu[d], Taomei Guo[d]*

[a]University of Chicago, [b]University of Massachusetts Amherst, [c]University of California Santa-Cruz, [d]Beijing Normal University

***Corresponding Authors:**

Ming Xiang
Linguistics Department
University of Chicago
mxiang@uchicago.edu

Taomei Guo
State Key Laboratory of Cognitive Neuroscience and Learning
Beijing Normal University China
guotm@bnu.edu.cn

Abstract

In wh-in-situ languages like Mandarin Chinese, for wh-questions, although the wh-phrase remains in a canonical argument position, syntactic theories generally posit that a covert dependency between the in-situ position and a clause-initial syntactic operator must nonetheless hold at the Logical Form, rendering wh-in-situ languages and wh-fronted languages abstractly similar. This paper investigates whether the processing of Mandarin wh-in-situ questions indeed involves constructing a long-distance dependency. Using the multiple-response speed-accuracy (SAT) paradigm. We show that Chinese wh-in-situ quesitons incur more processing costs compared to their non-wh counterparts. Furthermore, the length of the covert dependency only affected the processing accuracy, but not the processing speed, suggesting a content-addressable memory process underlying the construction of wh-in-situ dependencies, similar to overt long distance dependencies in English.

Keywords: wh-in-situ, Chinese, long distance dependencies, covert dependencies, memory mechanisms, processing complexity

1. Introduction

Natural language exhibits a large range of non-local dependencies, such as the relationship between a pronoun and its referent, a relative clause head noun and its base argument position, or an ellipsis site and its antecedent. To successfully parse non-local dependencies, the parser necessarily needs to encode the two ends of a dependency, and maintain the structural and semantic relations between the two elements throughout the parsing process. A well-studied case in both the theoretical and the psycholinguistic literature are wh-dependencies, as in "*which ring did Mary think John would like?"* In this example, the wh-phase *which ring* is interpreted as the theme of the verb *like,* but in surface structure the verb and the wh-element form a non-local dependency. When the parser encounters the wh-phrase during online comprehension, it immediately prioritizes establishing the dependency between the wh-phrase and an appropriate thematic role assigner – what Frazier dubbed the 'active filler strategy' (Frazier & Flores d'Arcais, 1989). The parser projects potential hosts for the wh-phrase, in accordance with the licensing conditions of the grammar and in advance of direct evidence for a moved constituent (e.g., Stowe 1986, Traxler & Pickering 1996, Wagers & Phillips 2009).

Parsing strategies such as the active filler strategy in the example above seem to be straightforward for overt wh-dependencies, whose dependent elements are readily apparent in the input. That is, comprehenders may readily detect a gapped argument and a dislocated filler using only superficial cues. However, it is well-known that wh-dependencies that do not seem to present a non-local relationship on the surface can be found in many languages, such as Hindi, Japanese or Korean. These are commonly referred to as wh-in-situ dependencies, because the wh-element occurs in the position where it is interpreted in (e.g., "*John likes what*?"). The linear order of elements seems to preclude the equivalent of an active filler strategy in these constructions, since there is no visible element in the dependency until the wh-phrase is encountered in its thematically-interpreted position. Instead, once the wh-phrase is encountered, the comprehender must identify its scope relationship with the other sentence operators. If doing so involves building a covert long-distance dependency with a clause-edge position, then wh-in-situ constructions might engage the same types of parsing procedures involved in building overt wh-dependencies. In particular the comprehender may need to retrieve or reactivate information about clausal heads it has encountered. In the present paper we attempt to determine whether processing in situ *wh*-phrases involves the retrieval of the non-local sentence context by focusing on the time course of comprehension in Mandarin.

In Chinese wh-questions, the wh-word stays in its canonical argument position, rather than being moved to the edge of the clause, as shown in (1) below:

(1) Xiaoming chi-le shenme?
Xiaoming eat-perf what
What did Xiaoming eat?

Analyses of Chinese wh-questions fall into two broad classes: the *LF movement* approach and the operator-binding approach. The LF-movement analysis (Huang 1982) proposes that wh-in-situ questions share the same kind of syntactic structures and operations as their "overt-movement" counterparts, but the movement operations that dislocate the *wh*-word its scope position happen at Logical Form (LF), rather than at surface structure. In contrast to the LF-movement proposal, the *operator binding* approach (Aoun & Li 1993; Tsai 1994) suggests that the wh-phrases in wh-in-situ languages are not targeted for movement at either LF or at surface structure. Instead, they get their quantificational force by being bound by an interrogative operator that takes scope over the entire sentence (see Cheng 2003 for a review of competing theoretical analyses). Although different in many aspects, these two approaches share the same insight that wh-in-situ questions involve a non-local dependency, just like wh-questions that have overt movement of the wh-word. This is shown schematically in (2). For the LF-movement approach, the non-local dependency is established between the covertly-moved wh-word and its base position. For the operator-binding approach, the dependency is built between the wh-operator at the highest [Spec, CP] position and the wh-word, which remains in its canonical position. Crucially, both accounts maintain that the syntactic structure of a wh-in-situ question is very different from a non-interrogative sentence, despite the fact that on the surface a wh-in-situ question has the same linear word order as a non-interrogative sentence. Specifically, an abstract long distance dependency must be established between the [Spec, CP] position and the position where the wh-word is interpreted. Thus on either view, the parser needs to retrieve the matrix [Spec, CP] position for further processing, whether for the purposes of positing an interrogative operator in that position, or for the purposes of retrieving the landing site for a covert displacement operation targeting the in-situ wh-word.

(2) [$_{CP}$ wh [….e….]]
OR
[$_{CP}$ Q-operator [….wh….]]

Covert dependencies at LF have been postulated to explain a range of different syntactic phenomena, including scope ambiguities and antecedent contained deletion. For this reason LF as an independent level of syntactic representation is a position commonly assumed by generative linguists. However, the processing reflex of this level of representation, which is by definition distinct from the surface form of the linguistic input, has only relatively recently begun to attract attention from researchers interested in issues of real-time sentence interpretation (see Chen & Hale 2010; Frazier 1999; Frazier & Clifton 2000; Lidz & Musolino 2002). The relationship between the parsing steps taken to resolve in situ *wh*-phrases versus fronted *wh*-phrases is potentially a valuable source of constraints on the operation of the parser. One the one hand, overt and covert dependencies plausibly have core abstract similarities in their syntactic and semantic representations. However, the distribution of surface cues to the dependency are quite different. In covert dependencies, there is often no direct perceptual evidence for the existence of a dependency before the critical *wh*-word. While it is true that Mandarin interrogative prosody distinguishes wh-questions from declaratives, the relevant prosodic cues have been argued to be most prominent in the sentence final position (e.g. Lin 2004) or are affected by focus (Liu and Xu 2005). In a reading task as we had in the current study, during incremental comprehension of a *wh*-question, it is unlikely that prosodic cues serve as early cues to the existence of a *wh*-dependency[1]. Thus Mandarin *wh*-in-situ questions are unlike overt *wh*-dependencies, where early perceptual/morphological evidence (e.g., the early presence of a *wh*-phrase), unambiguously signals to the parser to prepare for a dependency.

In the current paper we ask if wh-in-situ questions show evidence of non-local dependency formation, as is the case for wh-fronted questions, in spite of the fact that in-situ dependencies are established covertly in wh-in-situ dependencies. If so, then we would expect that the same kind of processing mechanisms that are deployed to realize overt dependencies should be deployed in the covert ones as well. To this end, in the rest of this paper, we will first introduce findings that concern how the parser establishes and maintains overt wh-dependencies in linguistic working memory, and then present the results from a speed-accuracy-tradeoff (SAT) study on Chinese wh-in-situ questions.

2. Processing overt long distance dependencies

2.1. Memory structures underlying overt long distance dependencies

[1] In reading, once the comprehender realizes the current input should be parsed as a *wh*-question (e.g., upon encountering the wh-word), it is possible that the parser may generate the relevant covert prosody. But even so, covert question prosody should not be available to the parser before the input has been assigned a question analysis.

Because long distance dependencies are constructed by integrating temporally and structurally distant linguistic material, they require the support of working memory resources for their completion (Foraker and McElree, 2011). For an English wh-question like "*which book did John think Mary likes*?", two robust findings about the process of relating the displaced wh-word to its gap site have been observed. As mentioned above, one is the "active filler strategy" (Stowe 1986; Frazier & D'Arcais 1989; Fodor 1995): while reading a sentence starting with a wh-word, the parser actively looks for a gap and establishes a long distance filler-gap relation whenever possible rather than waiting for enough information to decide the exact position of the gap. For instance, Stowe (1986) reported that a sentence like "*My brother wanted to know who my brother will bring* ***us*** *home to at Christmas*" generally creates more processing difficulty and hence longer reading time at the word "*us*", because comprehenders predictively interpreted the wh-filler "who" as the direct object. Upon encountering "*us"* in direct object position, it becomes evident that comprehenders pursued an incorrect syntactic parse, causing difficulty. There have been a number of explanations for this strategy. One is that discharging the filler relieves a working memory burden, on the assumption that maintaining a filler in working memory is costly (Wanner & Maratsos 1978 cf. McElree, Foraker & Dyer 2003, Wagers 2012). Forward prediction effects have also been found for relationships that are not strictly filler-gap relationships, such as pronominal cataphora in English (Kazanina et al. 2007) or clitic left dislocation dependencies in Spanish (Pablos 2006). Therefore, in a more general sense, unlicensed grammatical features or dependencies may be costly or lead to changes in parsing priorities (Gibson 2000, Aoshima, Phillips, & Weinberg, 2004).

The second robust finding in the literature is the distance effect. In order to establish a dependency, the parser needs to retrieve the previously-processed filler from working memory at the gap position. The distance between the filler and the gap affects this retrieval process of the filler, resulting in longer reading times and decreased acceptability for longer dependency lengths (Gibson 1998; Warren & Gibson 2002; Van Dyke & Lewis 2003; Lewis & Vasishth 2005).

The increased processing difficulty with increased length in wh-dependencies could potentially be driven by two different processes. In order to process a standard English type wh-question, the parser needs to retrieve the wh-filler at the gap position. One possible source of the length effect is a decline in the strength or integrity of the encoding of the filler over the course of the wh-dependency. This may be attributed to time-based decay: the greater distance between the filler and the gap, the more the memory representation of the filler may decay (Lewis & Vasishth 2005). Alternatively, introducing greater distance between the filler and its gap could lead to an increase in retrieval difficulty because doing so leads to the introduction of more encodings into memory. As a consequence of populating the workspace with more encodings, some of which may share features with the filler, there is likely to be greater interference at retrieval time (Anderson & Neely, 1996). Either time-based decay of the filler representation or retrieval interference could lower the rate of retrieval success, which would then be reflected in decreased accuracy for the reactivation and reintegration of the filler into the sentence at the gap site.

Another source of the length effect lies in the possibility of extra processing steps required to retrieve a distant filler. This rather different account could emerge if the retrieval of the filler involves an iterative search through the encodings in memory, guided by the dominance relations of the phrase structure (or some other prominence scale that gives rise to locality domains). In this account, hierarchically closer syntactic categories would be inspected before more distant ones. Consequently the greater the syntactic distance between the filler and the gap, the longer it will take to reach the filler. On this view the increased processing difficulty indexes the increased time needed to access the filler in the memory. If each step in the search process has some likelihood of failure, then increasing the number of steps in the process could also affect the accuracy of the search as well.

Teasing apart the effects of accuracy and speed is crucial for understanding the nature of the increased processing cost associated with longer dependencies. More generally, the question of retrieval speed is of particular interest because it speaks to basic issues of how language comprehension is supported and constrained by the memory architecture of the parser. Under one class of hypotheses, memory representations are accessed through a serial search process that necessarily accesses some irrelevant intermediate material before finding the target entry in memory (Sternberg 1966, 1975; McElree & Dosher 1993). If this is the case for language comprehension, then adding more distance between the retrieval site and the retrieved target will slow the processing speed. Another class of hypotheses suggests that linguistic material is retrieved in a direct-access, content-addressable manner (McElree, 2000), a form of associative memory (see Kohonen, 1989). These models assume that the linguistic features on the retrieval target (i.e., morpho-syntactic, semantic or pragmatic features) also index its location in memory, and thus those features can be used as retrieval cues to allow direct access to the target. Using such a memory access procedures obviates the need to consult the encodings of irrelevant, intermediate linguistic categories. The central prediction of this hypothesis is that increased distance has no effect on retrieval speed, although it might still affect retrieval accuracy.

2.2 Speed Accuracy Tradeoff (SAT)

Experimental measures that generate a single index of comprehension difficulty – measures like reading time or comprehension accuracy – potentially conflate speed and accuracy (Wickelgren 1977). Comprehenders may (implicitly) modulate their response behavior in ways that favor low accuracy and faster speeds, or they may instead prioritize slower speeds and higher accuracy. This impedes our ability to directly measure processing speed using simple reaction times, because slow responses may reflect slow processing, careful processing, or both. In the present case, we aim to distinguish hypotheses that make different predictions about the speed of memory access, and so we require a methodology that allows us to control for this characteristic tradeoff between processing speed and processing accuracy.

One technique to accomplish this is the response-signal technique, also known as the speed-accuracy tradeoff (SAT) procedure (Dosher 1976; McElree & Dosher 1993; McElree et al. 2003; Foraker and McElree, 2011). The SAT technique controls the trade-off between speed and accuracy by measuring the accuracy achieved in a task at a variety of deadlines. From these measurements, a function is estimated that that relates accuracy to elapsed processing time. Participants in an SAT experiment are trained to judge a property of a test stimulus, such as the well-formedness of a sentence, within a series of predefined 100-300 ms response windows following presentation of a response cue (a tone). The windows are chosen so that there will be periods when comprehenders are at chance, periods when their accuracy is changing rapidly as a function of response time, and periods in which accuracy has reached a final, asymptotic level. By calculating accuracy at each possible response point, it is possible to derive the full time-course of processing. Accuracy is measured in these experiments not simply as percent correct, but as d' (MacMillan & Creelman 2004). The major advantage of using d′ is that it corrects for response bias. For example, suppose in a grammaticality experiment a participant responds

‘YES’ to grammatical sentences and ‘NO’ to ungrammatical sentences. A good measure of accuracy in this experiment takes into account not only how often the participant says ‘YES’ to grammatical sentences but also how often she says ‘YES’ to ungrammatical sentences. The d′ measure does this by taking the difference between the proportion of grammatical sentences classified as grammatical, called the ‘hit’ rate, and the proportion of ungrammatical sentences mistakenly classified as grammatical, called the ‘false alarm’ rate.

Figure 1 illustrates two typical SAT functions. A good model of these data is given by a shifted, saturated exponential function, with the parameters λ, δ, and β (but see Ratcliff, 2006, for an alternative):

(3) $$d' = \lambda(1 - e^{-\beta(t-\delta)}),\ \mathrm{t} > \delta,$$

$$d' = 0\text{, otherwise}$$

The functions derived from an SAT experiment can thus usually be divided into 3 phases:

(4) a. A period of chance performance

In the model function, this period is captured by the SAT intercept parameter, δ (ms). Where $\mathrm{t} \leq \delta$, d' is defined as 0. The arrowheads along the *x*-axes in Figure 1 identify δ.

b. A period of increasing accuracy

In the model function, the rate of increase is captured by the SAT rate parameter, β (ms^{-1}). 1/β is sometimes called the function’s time constant.

c. An asymptotic period

In the model function, asymptotic accuracy is captured by the SAT parameter, λ (d').

The SAT asymptote parameter, λ, measures the overall probability of a certain response, such as correctly classifying a sentence. It provides an estimate of information quality in the sense that it indexes the best a participant can do, given an indefinite amount of time to deploy the relevant processes. Figure 1, Panel A, depicts a hypothetical case in which asymptotic accuracy varies between two conditions, while other parameters are held constant. The condition corresponding to the solid line reaches a higher level of asymptotic accuracy than the condition corresponding to the dotted line, even though they approach their respective asymptotic accuracies at the same rate.

Crucially, the asymptotic accuracy parameter can be separated from the rate parameter, β, and the intercept parameter, δ, which jointly measure processing speed. Figure 1, Panel B, depicts a hypothetical case in which asymptotic accuracy is constant, but rate differs between the two functions. The condition corresponding to the solid line reaches asymptotic accuracy faster than the condition corresponding to the dashed line.

By characterizing a process' SAT function in terms of its asymptote, rate and intercept, it is possible to tease apart the differential effects a particular manipulation may have on the quality of the representations involved (the process's accuracy) and the speed with which those representations are manipulated. For more in depth discussion, see Wickelgren (1977), Dosher (1979), Meyer et al. (1988), inter alia.

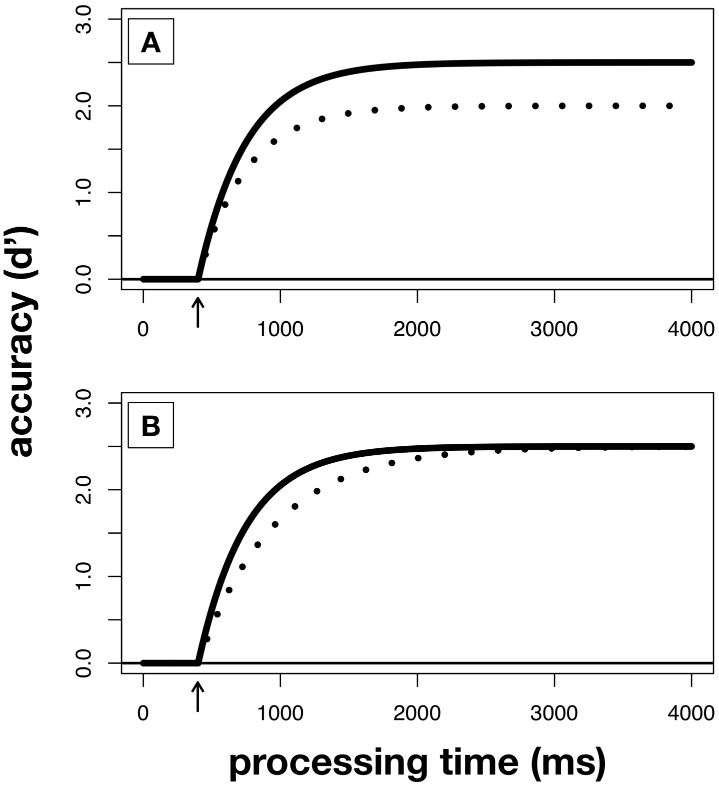

Figure 1: Two hypothetical SAT functions: Panel A depicts two functions that differ in asymptotic accuracy only, with rate and intercept held constant. Panel B depicts two functions that differ in rate only, with asymptote and intercept held constant. In both panels, the intercept is indicated along the *x*-axis by the arrow.

2.3. SAT and filler-gap dependencies

In distinguishing speed and accuracy, the SAT technique has been informative for the study of long-distance dependency formation in sentence comprehension. The crucial study has been McElree, Foraker & Dyer (2003), which examined the processing of filler-gap relations in the cleft sentences in (5). They were testing the hypothesis that comprehenders recovered the filler at the gap site by iteratively searching through the representations in memory. If so, they reasoned, the speed of dependency would vary with the distance between the filler and the gap. On the other hand, if access to the filler representation were direct, then the speed of dependency completion should not depend on the number of items in memory. They constructed stimuli that crossed two factors: thematic goodness-of-fit between a critical verb and its displaced argument; and the number of clauses separating the filler and gap:

(5) It was the scandal that …

(a) SAME CLAUSE: the celebrity relished/ *panicked.

(b) +1 CLAUSE: the model believed that the celebrity relished/*panicked.

(c) +2 CLAUSES: the model believed that the journalist reported that the celebrity relished/*panicked.

Participants judged the acceptability of these sentence types in the SAT paradigm. After deriving the SAT functions, McElree et al. (2003) found that a single rate parameter was sufficient to describe the functions for all three length conditions. This suggested that access to the filler representation did not depend on the amount of other linguistic material in memory. Accuracy, on the other hand, declined monotonically with dependency length. This suggested that longer dependencies were less likely to be successfully completed. This latter finding is consistent with a large body of data and theory showing a length effect on complexity (e.g., Gibson, 1998, 2000).

There are other explanations of the data that are consistent with the shared rate. The postulation of a special buffer for the filler, e.g., the HOLD cell of Wanner & Maratsos (1978), also predicts the patterns observed in the data, as long as the representation of the filler in this buffer is allowed to decay in strength. McElree et al. (2003) argued against the HOLD cell type account on the basis of other data. In particular, cross-modal lexical priming tasks (Swinney et al. 1988, McElree 2001) showed that lexical decision times for semantic associates of the filler were only facilitated at two points in the sentence: immediately after the filler's introduction at the clause-edge position and after the verb that presumably triggered its retrieval. But if the priming effects could be the consequence of the filler's decaying strength in the HOLD cell, then they do not provide especially diagnostic evidence. However, the HOLD cell account and the direct access mechanism can be reconciled if the filler were encoded with a special 'flag' that marks it as the head of an incomplete dependency. In a content-addressable memory, that 'flag' forms a part of the filler's address in a way that distinguishes it from all other encodings in the sentence, functionally implementing the HOLD cell (though not physically).

An incomplete dependency in an overt *wh*-movement language is signaled in the input in such a way that would allow a comprehender to give it a special encoding. For this reason, it is not an ideal test of the hypothesis that direct access is the characteristic memory access procedure comprehenders deploy. Long-distance dependencies in *wh*-in-situ languages provide a stronger test, since they do not announce their presence until the tail of the dependency is encountered, i.e. the wh phrase itself. If the comprehender were not expecting a *wh* dependency and thus not predictively encoding a clause-edge position for future retrieval, then encountering a *wh*-phrase might trigger an iterative search for an appropriate host. In other words, if searches are used in language comprehension, *wh*-in-situ constructions seem to provide an especially encouraging environment in which to use them.

3. The present study

3.1. Processing *wh*-in-situ questions

Content addressable access has been argued to be operative across a variety of dependencies, such as VP ellipsis (Martin & McElree, 2008) and subject-verb agreement (Wagers, Lau & Phillips, 2009). However, *wh*-in-situ questions differ from these non-local dependencies in at least one crucial aspect. For these other constructions, the items to be retrieved all carry some sort of semantic/pragmatic features (animacy, gender, etc) that can be used as a retrieval cue at the point of memory access. But if *wh*-in-situ questions are indeed processed as non-local dependencies, the target of retrieval is different from overt long distance dependencies. First, let's assume that the real time processing of *wh*-in-situ questions, just like overt dependencies, is also constrained by the limited focal attention, in the sense of McElree, Foraker & Dyer (2003). This means that some major aspects of the global syntactic and semantic representation constructed up to the current point is not immediately accessible for parsing operations. Information that was parsed earlier and beyond the current window of focal attention has to be retrieved back from working memory if they are needed. The target of retrieval for a *wh*-in-situ question at the *wh*-word is a purely syntactic position, either the [spec, CP] of the intermediate clause (in the case of an embedded wh-question) or the [spec, CP] of the matrix clause (in the case of a matrix question). At least for Mandarin Chinese, there is no overt marking of this position. It is not clear how pure structural positions of this kind are coded in linguistic working memory, and whether the parser targets these positions with the same sort of content-addressable access procedure in the absence of clear retrieval cues (which are not necessarily syntactic cues). Another important difference between the cleft examples examined by McElree et al. (2003) and wh-in-situ questions lies in the distribution of cues to the filler-gap dependency. For clefted filler-gap dependencies, the parser may engage anticipatory processes in advance of the gap once the filler has been recognized. As mentioned earlier, an overt filler could receive a special encoding from the parser, which flags the parser to anticipate a gap position. This could potentially explain the SAT data in McElree et al. (2003), without appealing to the distinction between a search and a content-addressable retrieval process. In the case of wh-in-situ questions, the parser isn't given any cue to expect a non-local dependency until it reaches the in-situ wh-word. Therefore the dependency for an in-situ question does not allow for anticipatory processing, and is constructed only when the in-situ *wh*-element is encountered. By hypothesis, this dependency construction involves the retrieval of the syntactic positions at which a question operator could be posited and then changing the representation to reflect the existence of the operator. For this reason, retrospective dependencies of this sort provide a more stringent test of the contrast between search and direct-access memory architectures.

A number of previous studies have examined the processing difficulty associated with in-situ *wh*-elements in Japanese. A *wh*-word in Japanese does not need to leave its canonical theta position, but it finds its scope position by associating itself with a corresponding scope marker attached to the verb at the appropriate position. Miyamoto and Takahashi (2003) investigated the processing of *wh*-elements that were inside of an embedded clause, and manipulated whether the scope marker *–ka* was found on the embedded or matrix verb. With a self-paced reading paradigm, they found that comprehenders showed longer reading times on non-scope-marked embedded verbs, relative to verbs in this position bearing a scope marker. Aoshima, Phillips and Weinberg (2004) extended and replicated these results. This finding is reminiscent of the 'active-filler' strategy for displaced wh-elements: upon encountering an in-situ *wh*-phrase in a head-final language like Japanese, comprehenders actively anticipate the presence of a scope-marker in grammatically licensed positions. In addition, it appears that the distance of the dependency between a *wh*-element and its scope position in Japanese has an impact on processing analogous to that observed in English. Ueno and Kluender (2009) found that longer wh-in-situ dependencies elicited a larger right-lateralized anterior negativity (RAN) compared the shorter ones (Expt2), suggesting more processing cost associated with longer in-situ dependencies

It is difficult however, to conclude from these results that in-situ dependencies involve the construction of a covert dependency. As in English, the dependency between the *wh*-element and its scope marker in Japanese is overt. If comprehenders adopt a prospective search for a scope marker upon encountering a *wh*-word, then the difficulty associated with increased length may be attributed to an in-situ analog of the familiar filled-gap effect. That is, comprehenders may eagerly anticipate a scope marker at the embedded verb position, which would prove to be a mistake if the scope marker were actually hosted by the matrix verb. If this were the case, then the Japanese in-situ length effects would reflect difficulty associated with revising an incorrect structural commitment. It remains to be seen whether in-situ dependencies that are not overtly marked (as in Mandarin Chinese) are processed in a similar way.

In the experiment we report below, we compared *wh*-in-situ questions with their declarative counterparts. In addition, we also manipulated the length of the *wh*-questions and declaratives. If processing *wh*-in-situ questions indeed involves the construction of a non-local dependency between the *wh*-element and an unobserved structural position, we expect either lower speed or accuracy for the questions compared to the control declaratives. In particular, we expect to see a length effect related to the retrieval of the [Spec,CP] position for *wh*-in-situ questions, but no such length effect for declaratives that do not require retrieval. Specifically, we predict a lower speed or accuracy for the longer in-situ questions compared to the shorter ones. A length effect above and beyond any complexity effect found with declarative sentences would lend support to the hypothesis that *wh*-in-situ dependencies are processed via the construction of a covert long-distance dependency.

If there is indeed a length effect for Mandarin in-situ *wh*-elements, then a secondary question concerns the source of this effect. The fact that Mandarin *wh*-dependencies cannot make use of anticipatory strategies makes them a good test case for asking about the processes that underlie the length effects. If the length effect is observed in slower processing speeds for long dependencies compared to short ones (either in terms of rate or intercept of the SAT function), then this would suggest a memory retrieval based on serial search, rather than a content addressable process. This finding would raise the possibility that the previous findings of constant retrieval speed for filler-gap dependencies reflects the effects of structural anticipation, rather than content-addressable retrieval. On the other hand, if the length effects are instead observed only in the asymptotic accuracy of the SAT function, then this would suggest a direct-access mechanism for retrieval of the [Spec,CP] position. This would provide strong evidence that the retrieval processes used to process *wh*-dependencies rely on direct-access mechanisms, and that the length effects observed in *wh*-dependencies reflect a diminished probability of successful retrieval for longer dependencies.

3.2. The current experiment

Participants

20 native Mandarin speakers from Beijing Normal University participated in the study for RMB50. All were university students between 22 and 25 years old. Each participant attended one 1-hour practice session and then six 1-hour experimental sessions. All of these sessions are separated by at least one day apart.

Material

For the critical conditions, 3 factors were crossed to create a 2×2×2 design. The factors were construction type (declarative or wh-question), acceptability (acceptable or anomalous), and length (short or long). An example is shown in (6a)-(6h). In the long wh-Q conditions, we only chose matrix verbs that could not take an interrogative complement (e.g. “order” in this example). This eliminated any ambiguity between a matrix and an embedded question, as every question must be interpreted as a matrix question. The semantically anomalous conditions were created by changing the last verb of the corresponding acceptable conditions to create a semantic mismatch between the verb and its object.

(6)

a/#b **Declarative; Short**
市政府 严惩了/#扩建了 那些官员。
shi zhengfu yancheng-le/kuojian-le nàxie guanyuan
city council punish/ expand those officials.
"The city council punished/expanded those officials."

c/#d **Declarative; Long**
市长 命令 市政府 严惩了/#扩建了 那些官员。
shizhang mingling shizhengfu yancheng-le/kuojian-le nàxie guanyuan
mayor order city council punish/expand those officials
"The mayor ordered the city council to punish/expand those officials."

e/#f **Wh-Q; Short**
市政府 严惩了/#扩建了 哪些官员?
shi zhengfu yancheng-le/kuojian-le nǎxie guanyuan
city council punish/expand which officials
"Which officials did the city council punish/expand?"

g/#h **Wh-Q; Long**
市长 命令 市政府 严惩了/#扩建了 哪些官员?
shizhang mingling shizhengfu yancheng-le/kuojian-le nǎxie guanyuan
mayor order city council punish/expand which officials
"Which officials did the mayor order the city council to punish/expand?"

In addition, we also included the unacceptable sentences in (6i) and (6j). These conditions were intended to control for the possibility that participants simply evaluate the local semantic (mis)match to perform the judgment task, rather than processing the whole sentence. These sentences contained multiple wh-dependencies (*why* and *which*), which rendered the sentences unacceptable[2]. In order to recognize the unacceptability, participants must have successfully constructed the *wh*-in-situ dependency and noticed the conflict with the highest *wh*-word. Furthermore, to prevent participants from establishing a simple low-level association between seeing the intital word ‘*weishenme (why)’* and judging the sentence to be unacceptable, we added acceptable why-questions that are maximally similar to (6i) and (6j), as shown in (6k) and (6l).

(6)(continued)

*i **Multiple-Wh; Short; unacceptable**
*为什么 市政府 严惩了 哪些官员?
Weishenme shizhengfu yancheng-le nǎxie guanyuan
why city council punish which officials
“Why did the city council punish which officials?”

*j **Multiple-Wh; Long; unacceptable**
*为什么 市长 命令 市政府 严惩了 哪些官员?
Weishenme shizhang mingling shizhengfu yancheng-le nǎxie guanyuan
why mayor order city council punish which officials
“Why did the mayor ask the city council to punish which officials?”

k. **Why-Q; Short; acceptable**
为什么 市政府 严惩了 那些官员?
Weishenme shizhengfu yancheng-le nàxie guanyuan
why city council punish those officials
“Why did the city council punish those officials?”

[2] It should be noted that there seems to be different types of multiple-wh questions in Chinese. Although wh-questions with multiple wh-arguments (e.g. subjects and object *wh*-phrases) are acceptable (Huang 1982), wh-questions involving the interaction between adjunct wh-phrases and argument wh-phrases seem to be degraded. Most relevant for the current discussion, Tsai (2008) observed that interrogative *weishenme* (why) does not form acceptable wh-questions with subject or object *wh*-arguments. The causal interrogative *zenme* (how) patterns identically. The acceptability judgments obtained in the current study (see below) confirm these intuitions.

l. **Why-Q; Long; acceptable**
为什么 市长 命令 市政府 严惩了 那些官员?
Weishenme shizhang mingling shizhengfu yancheng-le nàxie guanyuan
why mayor order city council punish those officials
Why did the mayor order the city council to punish those officials?

Thus for each item set there were 12 conditions (6a-l): 6 acceptable, and 6 unacceptable. A total of 40 such item sets were created. These 480 sentences were equally distributed into the 6 experimental sessions for each participants (see below). In addition to the 80 test sentences in each session, participants saw an additional 126 filler sentences in each session from unrelated experiments. Within each session, the order of the sentences was randomized, and the order of session was counterbalanced across participants.

Note that in previous SAT studies, the target of retrieval was also the target on which the lexical semantic features were manipulated to make the whole dependency either semantically plausible or implausible. This design ensures that the plausibility decision requires comprehenders to perform the target memory retrieval. In the current design, this was not the case: the semantic plausibility was evaluated locally between the verb and the in-situ *wh*-object, while the hypothesized retrieval of the *wh*-word's scope position was from a higher [spec, CP] position. If comprehenders judge the local semantic implausibility without calculating the *wh*-word's scope, then the resulting SAT curves would not be interpretable in light of our hypotheses. On the other hand, if calculating the *wh*-scope and establishing a long-distance dependency were carried out prior to any semantic evaluations, including local semantic (mis)match, the traditional SAT logic would still be applicable for the current design. Because it is not yet known whether in the current conditions participants perform the plausibility judgment task without fully processing the syntactic structure, the addition of the multiple *wh*-word conditions (i and j) are critical. When the acceptable *wh*-in-situ conditions (e and g) are scaled against these unacceptable multiple-*wh* conditions (i and j), the original SAT logic holds: unacceptability in these conditions can only be registered once the matrix scope of the *wh*-word has been computed. Thus, the resulting SAT curves unambiguously reflect at least the target retrieval process. By comparing the two *d'* scaling schemes (i.e. *e* and *g* scaled against *f* and *h;* and *e* and *g* scaled against *i* and *j*), we would know whether when performing the experimental task in *e* and *g*, participants have by-passed the stage of retrieving the target structural position.

Procedure

We employed the multiple-response SAT paradigm, following Wickelgren, Corbett & Dosher (1980). Stimulus presentation, timing, and response collection were controlled with the Linger software by Doug Rohde (available at http://tedlab.mit.edu/~dr/Linger/). Each trial began with a 500 ms fixation cross presented in the center of the screen. Each word appeared in the center of the screen for 400 ms, followed by 200 ms of blank screen. All words were presented using simplified Chinese characters, and the last word of each sentence was marked with a period (。) or with a question mark (?). At the onset of the final word, a series of 18 auditory response cues (50 ms, 1000 Hz tone) was initiated. The cues occurred every 350 ms, and the final word of the sentence remained on the screen. Participants were trained to initially respond by pressing both response keys simultaneously to indicate an undecided response, to switch their response to either the 'accept' or 'reject' key as soon as they could, and to modulate that response appropriately if their opinion of the sentence's acceptability changed during the response period. These instructions were modeled after other multiple-response SAT studies (Martin & McElree, 2008).

Data analysis

Mean percentage accuracies of subjects' judgments were calculated from the last three tone responses to check for the outliers. 5 subjects (out of 20) and 8 item sets (out of 40) were excluded from the data analysis because one or more conditions from these subjects or item sets had an accuracy rate lower than 40%.

In the next step we transformed the comprehension accuracy into d′ scores. There are a total of 6 contrasts that are of central interest, as listed below:

(7) The six relevant contrasts and their corresponding condition numbers

(i) **Short declaratives (1a vs. 1b):**

Declarative short acceptable vs. Declarative short unacceptable

(ii) **Long declaratives (1c vs. 1d):**

Declarative long acceptable vs. Declarative long unacceptable

(iii) **Short Wh-Qs (1e vs. 1f):**

Wh-Q short acceptable vs. Wh-Q short unacceptable semantic anomaly

(iv) **Long Wh-Qs (1g vs. 1h):**

Wh-Q long acceptable vs. Wh-Q long unacceptable semantic anomaly

(v) **Short multiple-Wh-Qs (1e vs. 1i):**

WH-Q short acceptable vs. Wh-Q short unacceptable multiple-wh anomaly

(vi) **Long multiple-Wh-Qs (1g vs. 1j):**

Wh-Q long acceptable vs. Wh-Q long unacceptable multiple-wh anomaly

For each pair, we scaled the acceptable sentences against their unacceptable counterparts by taking the difference between the z-transformed hit rate and the z-transformed false alarm rate (MacMillan & Creelman 2005). The hit rate refer to the proportion of "Yes" answers when the sentence is indeed acceptable, and "false alarms" to the proportion of "Yes" answers when the sentence is actually unacceptable. Transforming the raw scores into d′ scores adjusts for the potential response bias that participants may have. For this reason the discriminative d′ prime is a more reliable measure of the participants' ability to discriminate the acceptable and unacceptable sentences.

The shifted exponential function in (3) was then fit to the grand mean d′ scores at each of the 18 tone positions using an iterative hill-climbing algorithm (Reed 1976) that minimizes the squared deviances of predicted values from the observed data. A hierarchical model-testing scheme was used to determine whether conditions share values for a parameter (asymptote, rate or intercept), or they all need distinct values. Model fits range from a null model, in which all conditions share the same set of parameters (1λ-1β-1δ) to a completely saturated model, in which each condition has a different set of parameter values (6λ-6β-6δ). For model selection, we first assessed the fit quality of each model by calculating an adjusted R^2-statistic (see Appendix 1). Adjusted R^2 follows the conventional least-squares estimation (LSE) method that favors the parameter values that minimize the sum of squares errors between the observed value and the predicted value, weighted by the number of parameters in the model. A larger R^2 indicates a better fit to the empirical data. Adjusted-R^2 for all the models ranged from 0.903 to 0.997. For the top 15 models (all R^2>0.996) we calculated the AIC and the BIC scores (following Liu & Smith 2009). These two measures penalize complex models that postulate more parameters more strongly than does adjusted R^2, and hence guards us from overfitting; AIC has a bias to closer fits, and BIC has a bias against more parameters (Wagenmakers and Farrell 2004; Wagenmakers 2007). Smaller AIC and BIC values indicate a better fit. In addition, following Liu and Smith (2009), we provide another measure, the deviance, of each model by calculated by taking -2 times the log-likelihood of the data under a given model. The smaller the deviance, the better fit of the model (conversely, the greater the log-likelihood of the data, the better the model fit). The algorithms to calculate these additional measures are also included in Appendix 1.

In addition to the competitive model fits of the grand mean d′ scores, we also fitted each individual participant's SAT functions. The results of the individual subject fits were compared to check for consistency in the parameter estimates suggested by the grand mean model, ensuring that any patterns observed in the average d′ scores reflect patterns that were reliably observed in individual participants.

Results

The top 5 models, chosen based on their adjusted R^2, AIC and BIC scores, are presented in Table 1.

Table 1 The top 5 models with the best model fit

Model	adjusted R^2	deviance	AIC	BIC
4λ-2β-1δ	0.996052	-209	-195	-176*
4λ-3β-1δ	0.996163	-212*	-196*	-174
4λ-4β-1δ	0.996123	-212	-194	-169
4λ-5β-1δ	0.996187	-212	-192	-163
5λ-2β-1δ	0.996226*	-208	-192	-171

These five models differ very little in their adjusted R^2 scores, In light of the narrow range of variation between adjusted R^2 scores, it is informative to look instead at BIC and AIC scores, which penalize model complexity in different ways than does adjusted R^2. BIC more strongly penalizes more complex models than does AIC, and so to keep a good balance between model fit and model complexity we selected both 4λ-2β-1δ and 4λ-3β-1δ as the final candidates due to each being the best fit model on the BIC and AIC metrics, respectively. Put differently, these two models are the simplest models that provide reasonably good fits to the data.

The two winning models agree on the number of parameters necessary to model the λ(asymptote) and δ(intercept) of the SAT functions. Both models posit four λ (asymptote) parameters, and allocate one shared λ value for the first three contrasts (a/b, c/d, e/f, see 7) and a separate λ value for each of the remaining three contrasts (g/h, e/j, g/l). Likewise, both models agree that all contrasts are best fit with a single intercept parameter. The difference between the two winning models is that the 4λ-2β-1δ model assigns two β parameter values, one for the short and long declarative contrasts together (a/b and c/d), and the other for the remaining four wh-question contrasts. In contrast, the 4λ-3β-1δ model allocates one rate parameter β for the two declarative contrasts together (a/b, c/d), and a separate β for the short and long contrasts within *wh*-questions. That is, under the 4λ-3β-1δ model, short *wh*-questions (e/f and e/j) receive a different β value from long wh-questions (g/h and g/l).

Whether or not long and short *wh*-contrasts receive different rate parameters is of critical theoretical importance. To evaluate whether we indeed need 3 different β values, we fit both models to each individual subject's data and derived a β value for each individual participant's data. The averaged estimates based on each participant's individual data is included in Appendix 2. Paired-t tests showed that two β values derived through the 4λ-2β-1δ model are significantly different ($\beta_1 = 2.22 \pm 0.34\ s^{-1}$, $\beta_2 = 1.26 \pm 0.13\ s^{-1}$; t(14)=3.9, p<.01). As for the three β values derived from the 4λ-3β-1δ model ($\beta_1 = 2.21 \pm 0.34\ s^{-1}$; $\beta_2 = 1.25 \pm 0.13\ s^{-1}$; $\beta_2 = 1.30 \pm 0.13\ s^{-1}$), the β value for the short and long declarative contrasts is significantly different from that of the short *wh*-questions (β_1 v. β_2; t(14)=4.2, p<.01) and the long *wh*-questions (β_1 v. β_3; t(14)=3.2, p<.01); however, the β values for the short wh-questions is not different from the long wh-questions (β_2 v. β_3; t(14)=0.54, p>.5). This suggests that the rate distinction between short and long *wh*-contrasts in the 4λ-3β-1δ model does not reflect a reliable generalization over individual subjects' data, and may reflect overfitting of the empirical data. Hence we chose the 4λ-2β-1δ model as the final best fit. The estimated values for each parameter in the 4λ-2β-1δ model and their corresponding contrast are shown in Table 2.

Table 2 Parameter estimates for the final 4λ-2β-1δ model, based on parameters estimated over the average lag-latency function. Appendix 2 gives the average parameters estimated over individual participant functions. *d'* scaling using the multiple *wh*-question false alarm rate is presented in parentheses.

	Accuracy (λ, *d'*)	Speed (β, 1/s)	Speed (δ,s)
Short declarative	3.97	1.43	0.8
Long declarative	3.97	1.43	0.8
Short wh-Q	3.97 (3.77)	0.97	0.8
Long wh-Q	3.54 (3.34)	0.97	0.8

To visualize the effects more clearly, we plot the data and overlay it with the function given by the 4λ-2β-1δ model in two separate figures below. In Figure 2a (top panel) we plot the data and curves from two declarative contrasts and the two *wh*-conditions that are scaled against their corresponding semantic anomalous conditions. In Figure 2b (bottom panel) we repeat the two declarative conditions (again scaled against the semantic anomaly conditions) and plot the two *wh*-conditions that are scaled against the multiple wh-conditions.

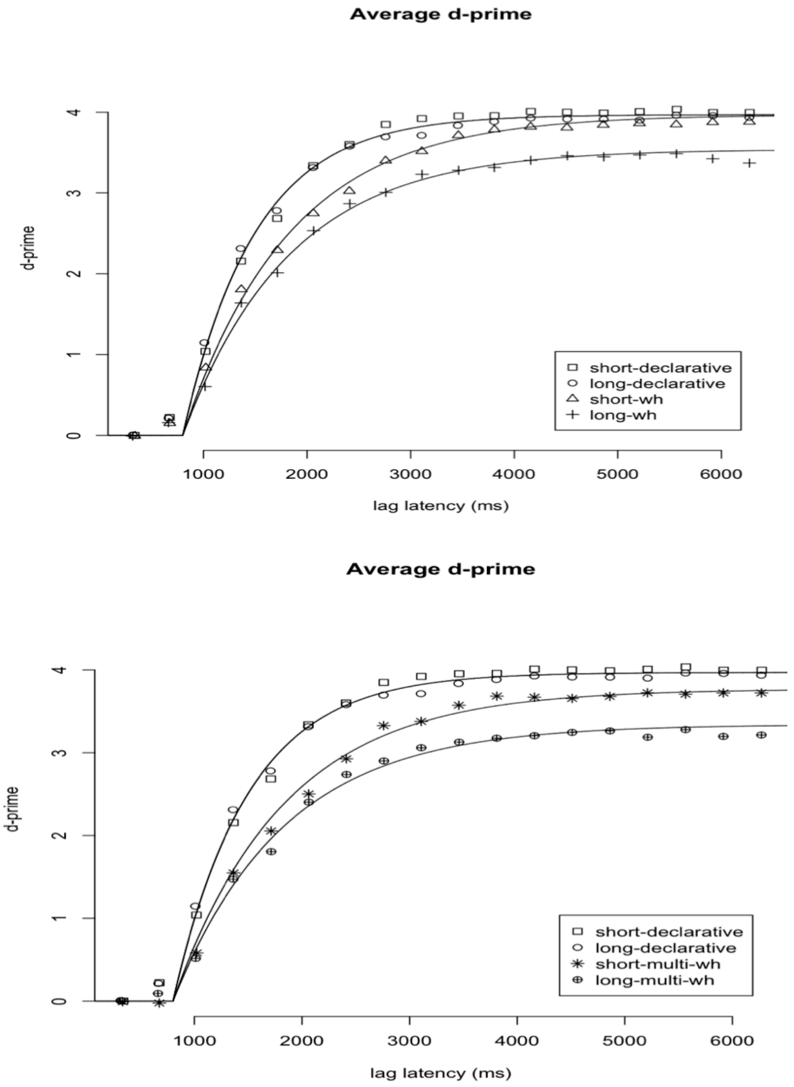

Figure 2. 2A (top panel): The two declarative contrasts and the two wh-contrasts with the semantic anomaly scaling. **2B** (bottom panel): the two declarative contrasts and the two wh-contrasts with the multiple-wh scaling.

As the best-fit model on the averaged data, the 4λ-2β-1δ model suggests that, first, there is a qualitative difference between declaratives and *wh*-in-situ questions; and second, the processing accuracy and the processing speed were affected differently by the length manipulation. For asymptote, the group estimate is presented in Table 2 below, and the averages over the individual estimates are presented in Appendix 2. The declarative conditions have larger asymptotes (group estimate 3.97 in d′ unit) than the wh-conditions (ranging from 3.34 to 3.97 in d′ unit). Paired t-tests over the individual estimates confirmed that the comprehension accuracy on declarative conditions is generally better than the wh-conditions (all ps<.05). More crucially, the short and long declarative conditions share the same asymptote (3.97), but the short wh-conditions have a larger asymptote than the longer wh-conditions: 3.97 vs. 3.54 when scaled against the semantically anomalous wh-conditions (p<.05); 3.77 vs. 3.34 (p<.05) when scaled against the multiple-wh conditions. We also looked at the actual empirical comprehension accuracy for each individual subject by averaging the d′ scores across the last three tone positions for each subject (Appendix 3). A 2×2 ANOVA on the two declarative contrasts and the two wh-contrasts with the semantic anomaly scaling showed a main effect of construction type F(1,14)=23.05, p<.001; a main effect of length F(1,14)=9.3, p<.01; and also an interaction between the two F(1,14)=6.5, p<.05. Paired comparisons showed no difference between the short and long declarative contrasts (t(14)=.875, p>.3), but a significant difference between the two wh-contrasts (t(14)=3.03, p<.01). Another 2×2 ANOVA on the two declarative contrasts and the two wh-contrasts with the multiple-wh scaling also found a main effect of construction type F(1,14)=22.98, p<.001; a main effect of length F(1,14)=14.06, p<.01; and an interaction between the two F(1,14)=8.99, p<.05. A pair comparison between the two wh-contrasts here showed a significant difference t(14)=3.8, p<.01.

For the rate parameter, the 4λ-2β-1δ model contained two distinct rates, one for the two declarative contrasts and the other for the other four wh-contrasts. The wh-conditions have a lower rate (0.97 s^{-1}) than the declarative conditions (1.43 s^{-1}), suggesting that *wh*-questions are processed more slowly than are correspondingly complex declarative sentences. One way to conceptualize this difference, which is perhaps more intuitive than the considering just rate parameter alone, is to compute the sum of the intercept with the inverse rate ($\delta + 1/\beta$). This figure gives the absolute time at which a characteristic percentage of asymptotic accuracy is achieved (~ 63%; or 1 – e^{-1}). Considering the parameters estimated over the average participant data (Table 2), declaratives attain this value 332 ms earlier than *wh*-questions. Considering the average of the parameters estimated over individual participants' data (Appendix 2), declaratives are 344 ms faster than *wh*-questions.

Critically, within the wh-contrasts, neither the length of the wh-questions nor the different scaling standard (semantic anomaly or multiple wh-questions) made a difference on processing rate; all wh-contrasts shared the same dynamics parameters. In addition, the length of the declarative conditions did not require separate intercepts. Thus the speed of processing declarative and interrogative sentences was unaffected by the length manipulation. All contrasts shared the same intercept under the 4λ-2β-1δ model.

4. Discussion

In this experiment we compared the processing dynamics of Chinese interrogative sentences containing a *wh*-in-situ phrase with declarative sentences, both at different lengths. We did this using the speed-accuracy tradeoff technique for decomposing the level of accuracy a participant is able to obtain in a task (asymptotic accuracy) from the speed at which that level is obtained (intercept, and rate). There were three major findings.

Firstly, we observed a significant difference in the processing speed of declarative contrasts in comparison to interrogative contrasts: interrogatives were processed more slowly than than declaratives. Secondly, declarative and interrogative contrasts were differentially affected by the length manipulation. Biclausal interrogatives attained a lower asymptotic accuracy than monoclausal interrogatives, even when the source of the anomaly constrast was an entirely string-/clause-local verb-theme pairing. In contrast, declarative sentences were unaffected by the length manipulation. Finally, and crucially, effects of length were limited only to asymptotic accuracy. Neither declarative nor interrogative contrasts revealed any difference in processing speed as a function of length. Importantly, these differences were observed with two *d'* scaling procedures: using the false alarm rate from multiple wh-questions and local semantic implausibility produced the same pattern of results. This provides strong converging evidence for the pattern of results observed here.

This pattern of results suggests that Chinese *wh*-in-situ questions are processed differently than their declarative counterparts, even though on the surface the two share identical word orders. In declarative constructions, DP arguments stand in a local relationship with the verb that selects for them. Thus it is perhaps unsurprising that overall length of the declarative sentence did not affect the local integration of a DP argument into the local verb. This was reflected in the finding that neither the speed of retrieval nor the overall acceptability judgment accuracy was affected by the length manipulation within the declarative contrasts. In contrast to this, *wh*-in-situ questions did show sensitivity to the length manipulation. Although the wh-words superficially appear to engage in a local relation with the selecting verb, their processing profile is more similar to that observed for overtly non-local dependencies. Crucially, *wh*-in-situ questions displayed clear length effects, with longer dependencies giving rise to lower accuracy in the SAT function. This mirrors findings about the effect of the length of overt dependencies on acceptability, lending support to the claim that *wh*-in-situ elements involve a covert dependency with their scope position (here, the matrix [Spec,CP]). This is compatible with either covert LF-movement analyses (Huang 1982), or unselective binding approaches (Aoun & Li 1993; Tsai 1994) to the *wh*-in-situ phenomenon. In either case, processing the wh-element involves identifying its scope position and associating the *wh*-element with that position. A second interesting data point that supports this conclusion is that *wh*-in-situ contrasts revealed a slower retrieval speed (as measured in the rate parameter) than the declarative contrasts. This processing speed difference between interrogatives and declaratives may reflect the extra parsing and retrieval processes necessary to associate the *wh*-in-situ element with its scope position. With respect to the first major theoretical question we set out to answer, these results show that *wh*-in-situ involves the processing of a covert dependency.

An additional theoretical question of interest concerned the memory access procedures necessary to parse such a covert dependency. By hypothesis, parsing the *wh*-in-situ required the retrieval of the scope position of the *wh*-element. However, the speed with which this retrieval occurred was of crucial interest. In a memory architecture that requires serial search, the length of the *wh*-dependency should result in longer retrieval times, on the assumption that the processor must first consult an additional (and unusable position) in the case of biclausal dependencies. In contrast, a content-addressable architecture predicts invariant retrieval speed for long and short dependencies, because in this architecture the speed of retrieval is set-size invariant: identifying the target matrix [Spec, CP] position does not depend on the number of already-encoded intermediate [Spec,CP] positions. We found that the rate parameter of the SAT function was constant across long and short *wh*-dependencies. This finding held true whether the acceptable wh-questions were scaled against unacceptable ones with a local semantic anomaly, or scaled against multiple-wh questions in which the anomaly stems from a non-local relationship. This finding suggests that Chinese speakers are able to access the matrix [Spec,CP] position without being slowed by the existence of intermediate positions, confirming a key prediction of content-addressable memory architectures. Previous research using the same methodology obtained similar results for English (McElree, et al., 2003). However, English *wh*-

phrases are dislocated to the left, potentially providing an anticipatory cue and consequently obscuring the existence of a search or other set-size dependent process. The fact that Chinese speakers could not be guided by any surface distributional cues to anticipate a long-distance dependency provides important convergent evidence that the earlier conclusions about retrieval mechanisms in filler-gap dependencies may hold independent of any "look-ahead" mechanisms.

There is one potential objection to our claim that processing Chinese *wh*-in-situ questions indeed involves establishing a non-local relationship that is absent in declarative sentences. It is possible that the observed effects on *wh*-questions arise from reanalysis due to "garden-path" effects. Since there is no explicit marker to indicate to the parser the current structure is a *wh*-question until the sentence-final *wh*-element is encountered, participants might have committed themselves to a declarative structure at an earlier point in the parse. If this occurred, then upon reaching the final *wh*-element, the comprehenders would need to undo the existing interpretation and establish a *wh*-construction. If this reanalysis procedure is sensitive to the complexity or length of the sentence, then the selective length-sensitivity observed for *wh*-contrasts might be expected. However, comparisons to both the multiple *wh*-questions and the local semantic anomalies revealed an identical pattern of results. In the case of the multiple *wh*-questions, the presence of an early *wh*-operator allows comprehenders to avoid any garden path that might result from maintaining a commitment to a declarative interpretation. For this reason, the present results are not likely to reflect an unexpected and difficult shift from declarative to interrogative force. A related possibility is that comprehenders maintain expectations about the likelihood of an upcoming *wh*-word, and that these expectations differ across our long and short conditions. If this were so, then the observed difference in accuracy might reflect these different expectations, rather than difficulty of retrieval. This account is unlikely, however, in light of a recent study by

Xiang, Liu, Chen and Guo (2011). These authors collected ERP data that compared matrix *wh*-in-situ questions and embedded ones (as in *Laoshi wen Yuehan baokao-le na-suo-xuexiao ("the teacher asked which school John applied to."*)). In addition to manipulating the scope height, they also manipulated the length of the dependency (short vs. long). Embedded *wh*-in-situ questions involve a non-local dependency between the *wh*-word and an implicit question operator at the embedded [Spec, CP] position. Since this operator is not phonetically marked, the parser needs to establish a covert *wh*-dependency as in the matrix questions. However, unlike the matrix *wh*-in-situ questions, in embedded questions the matrix verb (e.g. *wen (ask)*) obligatorily takes an interrogative as its complement[3]. The interrogative complement could either be a yes-no question or a *wh*-question. Speaker intuition doesn't suggest a strong bias for either possibility. After encountering the matrix verb that takes an interrogative complement, speakers' expectations for a wh-element should be greater than in a matrix-question scenario in which the matrix verb did not signal any upcoming interrogative. If speakers maintained expectations of varying strength about the likelihood of encountering a wh-element, one predicts differential processing costs for matrix questions and embedded questions. However, this wasn't borne out in Xiang et al. (2011). ERP results showed that both the matrix and embedded *wh*-in-situ questions elicited a positivity compared to the declarative controls. Crucially, however, the two wh-conditions didn't differ in the amplitude of their positivity. This suggests that comprehenders did not experience differential processing difficulty as a function of their expectations about upcoming *wh*-words. In light of this result, it seems unlikely that the present SAT results are

[3] Although in Mandarin not all complement-clause taking verbs that obligatorily require an interrogative as its complement, Xiang et al (2011) particularly chose verbs like "ask", which unambiguously signal an upcoming interrogative complement.

significantly driven by comprehenders' expectations about the likelihood of upcoming *wh*-elements.

5. Conclusions

Using the speed-accuracy trade-off paradigm, we demonstrated that interpreting a *wh*-in-situ question indeed evokes greater processing effort than a corresponding declarative. Both the speed and accuracy of interpretation were impacted. The processing profile revealed here is very similar to what has been observed for overt long-distance dependencies, lending support to the hypothesis that a covert long distance dependency is being constructed when Chinese speakers process *wh*-in-situ questions. We also argue that the underlying memory structure that supports the processing of a covert *wh*-dependency is content-addressable: the parser can directly target the [Spec, CP] position without searching through intermediate positions, despite of the fact that this is purely a syntactic position without any overt phonetic or semantic content.

Acknowledgments:
References:

Anderson, M. and Neely, J. 1996. Interference and inhibition in memory retrieval. *Memory*. Academic Press Inc. 237-313.

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

Aoshima, S., Phillips, C., and Weinberg, A. 2004. Processing filler-gap dependencies in a head-final language. *Journal of Memory and Language*, 51, 1: 23-54.

Chen, Z., & Hale, J.T. 2010. Deforesting Logical Form. In Ebert, C., Jäger, G, & Michaelis, J. (eds.). *The Mathematics of Language. Lecture notes in Artificial Intelligence*, 6419, 13-28, Springer: Heidelberg.

Cheng, L.L.S. 2003. Wh-in-situ. *Glot International* 7, 4: 103-109

Dosher, A. 1976. The retrieval of sentences from memory: A speed-accuracy study. *Cognitive Psychology* 8, 3: 291-310.
Fodor, J. D. 1978. Parsing strategies and constraints on transformations. *Linguistic Inquiry*, 9: 427-473.

Frazier, L. & Flores d'Arcais, G.B. 1989. Filler-driven parsing: A study of gap filling in Dutch. *Journal of Memory and Language*, 28: 331-344.

Frazier L. 1999. *On Sentence Interpretation*. Dordrecht, The Netherlands: Kluwer

Frazier L. and Clifton C., Jr. 2000. On bound variable interpretations: The LF-only hypothesis. *Journal of Psycholinguistic Research*, 29:125–140.

Foraker, S. and McElree, B. 2011. Comprehension of linguistic dependencies: Speed-accuracy tradeoff evidence for direct-access retrieval from memory. *Language and Linguistics Compass, 5*, 11: 764-783.

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. In Y. Miyashita, P. Marantx & W. O'Neil (Eds.), *Image, Language, Brain* (pp. 95-112). Cambridge, MA: MIT Press.

Huang, C.-T. J. 1982. Logical Relations in Chinese and the Theory of Grammar, Ph.D. dissertation, MIT.

Kazanina, N., Lau, E., Lieberman, M., Yoshida, M. and Phillips, C. 2007. The effect of syntactic constraints on the processing of backwards anaphora. *Journal of Memory and Language* 56, 3: 384-409

Kohonen, T. 1977. Associative memory, a system theoretical approach. New York, SpringerVerlag.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. 2006. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Science*, 10, 10: 447-454

Lidz,J. and Musolino, J. 2002. The scope of isomorphism: turning adults into children. *Language Acquisition*, 11, 4: 277-291.

Lin, M. 2004. On production and perception of boundary tone in Chinese intonation. Proc. Int. Symp. Tonal Aspects Languages: with Emphasis on Tone Languages, Beijing, pp. 125–129

Liu, F. and Xu, Y. 2005. Parallel Encoding of Focus and Interrogative Meaning in Mandarin Intonation. *Phonetica*, 62: 70-87

Liu, C-C and Smith, P-L 2009. Comparing time-accuracy curves: beyond goodness-of-fit measures. *Psychonomic Bulletin and Reviews*, 16, 1: 190-203

Macmillan NA, Creelman CD. *Detection theory: A user's guide. New York*: Cambridge University Press; 2004.

Martin, A. and McElree, M. 2008. A content-addressable pointer mechanism underlies comprehension of verb-phrase ellilpsis. *Journal of Memory and Language*, 58: 879-906

Meyer, D. E., Irwin, D. E., Osman, A. M., & Kounois, J. 1988. The dynamics of cognition and action: Mental processes inferred from speed-accuracy decomposition. *Psychological Review*, 95: 183–237.

McElree, B. & Dosher, B.A. 1993. Serial retrieval processes in the recovery of order information. *Journal of Experimental Psychology: General*, 122: 291-315.

McElree, B. 2000. Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research*, 29: 111–123.

McElree, B., Foraker, S. & Dyer, L. 2003. Memory structures that subserve sentence comprehension. *Journal of Memory and Language*, 48: 67-91.

Miyamoto, E. T., & Takahashi, S. 2000. The processing of wh-phrases and interrogative complementizers in Japanese. In N. Akatuka & S. Strauss (Eds.), *Japanese/Korean linguistics,*10: 62–75. Stanford, CA: CSLI.

Pablos, L. 2006. Pre-verbal Structure Building in Romance Languages and Basque. College

Park, University of Maryland Dissertation.

Ratcliff, R. 2006. Modeling response signal and response time data. *Cognitive Psychology*, 53: 195-237.

Reed, A. V. 1976. The time course of recognition in human memory. *Memory & Cognition*, 4: 16–30.

Sternberg, S. 1966. High speed scanning in human memory. *Science,* 153: 652–654.

Sternberg, S. 1975. Memory scanning: new findings and current controversies. *Quarterly Journal of Experimental Psychology*, 27: 1-32.

Stowe, L.A., 1986. Parsing WH-constructions: evidence for on-line gap location. *Language and Cognitive Processes,* 1: 227–245.

Swinney, D., Ford, M., Frauenfelder, U., & Bresnah, J. 1988. On the temporal course of gap-filling and antecedent assignment during sentence comprehension. In B. Grosz, R. Kaplan. M. Macken. & 1. Sag (Eds.), *Language structure and processing,* Stanford, CA: CSLI.

Traxler, M. J., & Pickering, M. J. 1996. Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, 35: 454–475.

Tsai, W.-T. D. 1994. On Economizing the Theory of A-bar Dependencies. Ph.D. dissertation. MIT.

Ueno, M. & Kluender R. 2009. On the processing of Japanese wh-questions: An ERP study. *Brain Research*, 1290: 63-90.

Wagenmakers, E-J and Farrell, S. 2004. AIC model selection using Akaike weights. *Psychonomic Bulletin and Review*. 11, 1: 192-196.

Wagenmakers, E.-J. 2007. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14: 779-804.

Wagers, M., Lau,E. and Phillips, C. 2009. Agreement attraction in comprehension: representations and processes. *Journal of Memory and Language*, 61: 206-237

Wagers, M. and Phillips, C. 2009. Multiple dependencies and the role of the grammar in real-time comprehension. *Journal of Linguistics*, 45: 395-433.

Wagers, M. 2012. Memory mechanisms for wh-dependency formation and their implications for islandhood. In J. Sprouse and N. Hornstein (Eds.) *Experimental Syntax and Island Effect*. Cambridge.

Wanner, E., & Maratsos, M. P. 1978. An ATN approach to comprehension. In M. Halle, J. Bresnan, & G. A. Miller (Eds.), *Linguistic Theory and Psychological Reality*. Cambridge, MA: MIT Press.

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

Wickelgren, W. 1977. Speed-accuracy tradeoff? and information processing dynamics. *Acta Psychologica*, 41: 67–85.

Wickelgren, W. A., Corbett, A. T., & Dosher, B. A. 1980. Priming and retrieval from short- term memory: A speed-accuracy tradeoff analysis. *Journal of Verbal Learning and Verbal Behavior*, 19: 387–404.

Van Dyke, J. A., & Lewis, R. L. 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, 3: 285-316.

Xiang, M.; Liu, FQ; Chen, PY; and Guo,TM. 2011. Cross-linguistic variations and similarities: an ERP study of Mandarin wh-constructions. Poster presented at the 24th CUNY Conference on Human Sentence Processing, Stanford.

Appendix 1

(a) $$adjusted\ R^2 = 1 - \frac{\sum_{i=1}^{n}(d_i - \hat{d}_i)^2/(n-k)}{\sum_{i=1}^{n}(d_i - \bar{d})^2/(n-1)}$$

(b) $$\mathrm{AIC} = -2\log L(\hat{\theta}) + 2k$$

(c) $$\mathrm{BIC} = -2\log L(\hat{\theta}) + k\log n$$

(d) $$L(\theta) = \prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(d_i - \hat{d}_i)^2}{2\sigma^2}\right]$$

* d_i is the observed d-prime value; $\hat{d}_i$ is the predicted d-prime; $\bar{d}$ is the mean; n is the number of data points; k is the number of parameters; $L(\theta)$ is the likelihood of the data given a fully specified model, and $L(\hat{\theta})$ is the maximized likelihood.

Appendix 2: The average estimates for the two competing models over each individual participant's data, with standard errors included in the parenthesis.

	4λ-2β-1δ			4λ-3β-1δ		
	Asymptote (λ) (d')	Rate (β) (1/s)	Intercept (δ) (s)	Asymptote (λ) (d')	Rate (β) (1/s)	Intercept (δ) (s)
Short declaratives (a vs. b)	λ_1=3.98 (0.06)	β_1=2.22 (0.34)	δ=0.93 (0.08)	λ_1=3.99 (0.06)	β_1=2.21 (0.34)	δ=0.93 (0.08)
Long declaratives (c vs. d)						
Short wh-Qs (e vs. f)		β_2=1.26 (0.13)			β_2=1.25 (0.13)	
Short multi-Wh-Qs (e vs. i)	λ_2=3.53 (0.17)			λ_2=3.53 (0.17)		
Long wh-Qs (g vs. h)	λ_3=3.77 (0.07)			λ_3=3.78 (0.08)	β_3=1.30 (0.13)	
Long multi- wh-Qs (g vs. j)	λ_4=3.32 (0.17)			λ_4=3.32 (0.16)		

Appendix 3: The actual empirical comprehension d-prime scores (in d'), with standard errors included in the parenthesis.

Short declaratives (a vs. b)	Long declaratives (c vs. d)	Short Wh-Qs (e vs. f)	Short multi-wh-Qs (e vs. i)	Long wh-Qs (g vs. h)	Long multi-wh-Qs (g vs. j)
4.0 (0.08)	3.94 (0.09)	3.88 (0.07)	3.37 (0.17)	3.72 (0.10)	3.21 (0.15)