

Processing covert dependencies: an SAT study on Mandarin *wh*-in-situ questions

Ming Xiang · Brian Dillon · Matt Wagers ·
Fengqin Liu · Taomei Guo

Received: 17 October 2011 / Accepted: 25 February 2013
© Springer Science+Business Media Dordrecht 2013

Abstract In *wh*-in-situ languages like Mandarin Chinese, in which the *wh*-phrase remains in a canonical argument position in *wh*-questions, syntactic theories generally posit that a covert dependency between the in-situ position and a clause-initial syntactic operator must nonetheless hold at logical form. *Wh*-in-situ languages and *wh*-fronted languages are in this way 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 tradeoff (SAT) paradigm, we show that Chinese *wh*-in-situ questions incur more processing costs than their non-*wh* counterparts. Furthermore, the length of the covert dependency affects only processing accuracy, but not processing speed. This pattern suggests 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

M. Xiang (✉)
Linguistics Department, University of Chicago, Chicago, IL, USA
e-mail: mxiang@uchicago.edu

B. Dillon
University of Massachusetts Amherst, Amherst, MA, USA

M. Wagers
University of California Santa Cruz, Santa Cruz, MA, USA

F. Liu · T. Guo (✉)
State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University,
Beijing, China
e-mail: guotm@bnu.edu.cn

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 to 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 is *wh*-dependency, as in “*Which ring did Mary think John would like?*” In this example, the *wh*-phrase *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 prioritizes establishing the dependency between the *wh*-phrase and an appropriate thematic role assigner—what Frazier dubbed the ‘active filler strategy’ (Frazier and 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 and Pickering 1996; Wagers and Phillips 2009).

The effects of parsing strategies like the active filler strategy 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, and 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*-phrase 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*-phrase to 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 and 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*-element, 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*-element is interpreted. Thus on either view, the parser needs to retrieve the appropriate [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*-element.

- (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 and Hale 2010; Frazier 1999; Frazier and Clifton 2000; Lidz and 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. On 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*-phrase. 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 like our current study, it is unlikely that prosodic cues serve as early cues to the existence of a *wh*-dependency during incremental comprehension of a *wh*-question.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 *wh*-in-situ dependencies are established covertly. 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 test this prediction, we conducted an experiment using the speed–accuracy tradeoff (SAT) technique (McElree et al. 2003) to measure behavioral responses of Chinese readers to *wh*-in-situ questions and to see whether we could detect the signatures of non-local dependency formation. Before we report the results, we first review findings concerning how the parser establishes and maintains overt *wh*-dependencies in linguistic working memory.

2 Processing overt long distance dependencies

2.1 Memory structures underlying overt long distance dependencies

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 and Flores d’Arcais 1989; Fodor 1995): while reading a sentence starting with a *wh*-phrase, 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

¹ In reading, once the comprehender realizes the current input should be parsed as a *wh*-question (e.g., upon encountering the *wh*-phrase), 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.

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 and Maratsos 1978; cf. McElree et al. 2003; Wagers 2013). 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 et al. 2004).

The second robust finding in the literature on *wh*-dependencies 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 and Gibson 2002; Van Dyke and Lewis 2003; Lewis and 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 the distance between the filler and the gap, the more the memory representation of the filler may decay (Lewis and Vasishth 2005; Lewis et al. 2006). 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 and 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 (cf. McElree et al. 2003). This rather different explanation 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 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 and Doshier 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 1977, 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 procedure 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 SAT procedure (Doshier 1976; McElree and Doshier 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 relates accuracy to elapsed processing time. Participants in an SAT experiment are trained to categorize a test stimulus by some property, 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 and 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 (after the proportions are first transformed to z-scores by the normal quantile function).

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):

$$(3) \quad d' = \lambda(1 - e^{-\beta(t-\delta)}), \quad t > \delta, \\ d' = 0, \quad \text{otherwise}$$

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

(4) a. *A period of chance performance*

In the model function, this period is captured by the SAT intercept parameter, δ (ms), where $t \leq \delta$, 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}). $1/\beta$ 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.

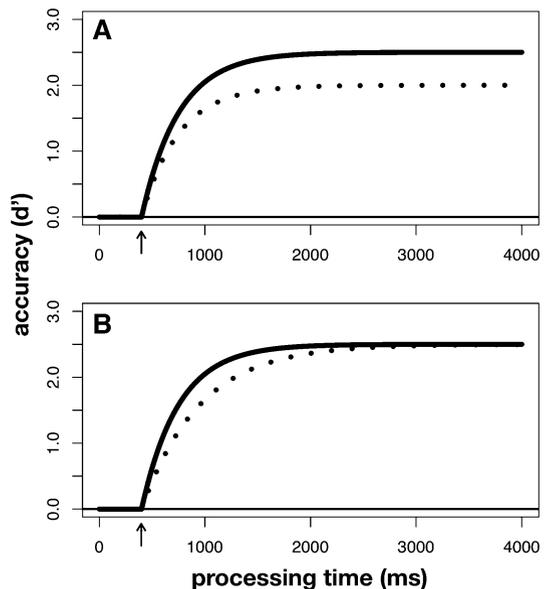
By characterizing a process's SAT function in terms of its asymptote, rate, and intercept, it is possible to tease apart the differential effects a particular manipulation may have on the quality of the representations involved (the process's accuracy) and the speed with which those representations are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

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 et al. (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.

Fig. 1 Two hypothetical SAT functions: **a** Depicts two functions that differ in asymptotic accuracy only, with rate and intercept held constant. **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 an arrow



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 and 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. One could continue to maintain the *HOLD* cell hypothesis if decline of the priming effects were attributable to the filler's decaying strength in the *HOLD* cell. However, perhaps a more attractive reconciliation of the *HOLD* cell account and the direct access mechanism is possible, if the filler were encoded with a special 'flag'—or unlicensed grammatical feature—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 and McElree 2008) and subject–verb

agreement (Wagers et al. 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 us assume that the real time processing of *wh*-in-situ questions, just like overt dependencies, is also constrained by a limited focal attention, in the sense of McElree et al. (2003). This means that, at any given point in a sentence, there are some features or constituents of the syntactic and semantic representation that will not be immediately accessible for parsing operations. Information beyond the current window of focal attention has to be retrieved or reactivated if it is needed. The target of retrieval for a *wh*-in-situ question at the *wh*-phrase 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 morphologically-unrealized syntactic categories 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*-phrase. 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*-phrase 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 (2000) investigated the processing of *wh*-elements that were inside 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 et al. (2004) extended and replicated these results. The 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 to shorter ones (their Experiment 2), 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*-phrase, then the difficulty associated with increased length may be attributed to the delayed resolution of an expectation for the scope marker. 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

3.2.1 Participants

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

3.2.2 Material

For the critical conditions, 3 factors were crossed to create a $2 \times 2 \times 2$ design. The factors were construction type (declarative or *wh*-question), acceptability (acceptable or anomalous), and length (short or long). Examples are shown in (6a)–(6h). In the long *wh*-Q conditions, we chose only 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
<i>city council</i>	<i>punish/ expand</i>	<i>those officials.</i>

‘The city council punished/#expanded those officials.’

c/#d Declarative; Long

市长	命令	市政府	严惩了/#扩建了
shizhang	mingling	shizhengfu	yancheng-le/kuojian-le
<i>mayor</i>	<i>order</i>	<i>city council</i>	<i>punish/expand</i>

那些官员。
nàxie guanyuan
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/ nǎxie guanyuan
 kuojian-le
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 would 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*-element. Furthermore, to prevent participants from establishing a simple low-level association between seeing the initial 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?’

² It should be noted that there seem to be different types of multiple *wh*-questions in Chinese. Although *wh*-questions with multiple *wh*-arguments (e.g., subject 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.

*j **Multiple-Wh; Long; unacceptable**

*为什么	市长	命令	市政府
*Weishenme	shizhang	mingling	shizhengfu
<i>why</i>	<i>mayor</i>	<i>order</i>	<i>city council</i>
严惩了	哪些官员?		
yancheng-le	nàxie guanyuan		
<i>punish</i>	<i>which officials</i>		

‘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
<i>why</i>	<i>city council</i>	<i>punish</i>	<i>those officials</i>

‘Why did the city council punish those officials?’

l. **Why-Q; Long; acceptable**

为什么	市长	命令	市政府
Weishenme	shizhang	mingling	shizhengfu
<i>why</i>	<i>mayor</i>	<i>order</i>	<i>city council</i>
严惩了	哪些官员?		
yancheng-le	nàxie guanyuan		
<i>punish</i>	<i>those officials</i>		

‘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 participant (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 the sessions 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*-phrase’s scope position was from a higher [Spec, CP] position. If comprehenders judge the local semantic implausibility without calculating the *wh*-phrase’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)matches, 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) was critical. When the acceptable *wh*-in-situ conditions (e and g) are scaled against these unacceptable multiple-*wh* conditions (i and j)—that is to say, when accurate performance in conditions e/g are taken as the hit rate and inaccurate performance in conditions i/j as the false alarm rate in the d' calculation—then the original SAT logic holds: Unacceptability in these conditions can be registered only once the matrix scope of the *wh*-phrase 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 tasks in e and g, participants have by-passed the stage of retrieving the target structural position.

3.2.3 Procedure

We employed the multiple-response SAT paradigm, following Wickelgren et al. (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, 1,000 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 and McElree 2008).

3.2.4 Data analysis

Mean percentage accuracies of subjects' judgments were calculated from the last three tone responses to check for the outliers. Five subjects (out of 20) and eight 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 and Creelman 2004). The hit rate refers 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 by offsetting the credit received for correct performance in an ‘expected yes’ condition with a penalty for incorrect performance in an ‘expected no’ condition.

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, λ ; rate, β ; or intercept, δ), or they all need distinct values. Model fits range from a null model, in which all contrasts share the same set of parameters ($1\lambda-1\beta-1\delta$) to a completely saturated model, in which each of the six contrasts in (7) has a different set of parameter values ($6\lambda-6\beta-6\delta$). 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 and Smith 2009). These two measures more strongly penalize complex models that postulate more parameters 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 calculated by doubling the negative log-likelihood of the model parameters given the data: The smaller the deviance, the

better fit of the model. The formulae 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.

3.2.5 Results

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

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 $4\lambda-2\beta-1\delta$ and $4\lambda-3\beta-1\delta$ 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) above) 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 $4\lambda-2\beta-1\delta$ 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 $4\lambda-3\beta-1\delta$ 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 $4\lambda-3\beta-1\delta$ 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

Table 1 The top five models with the best model fit

Model	Adjusted R^2	Deviance	AIC	BIC
$4\lambda-2\beta-1\delta$	0.996052	-209	-195	-176*
$4\lambda-3\beta-1\delta$	0.996163	-212*	-196*	-174
$4\lambda-4\beta-1\delta$	0.996123	-212	-194	-169
$4\lambda-5\beta-1\delta$	0.996187	-212	-192	-163
$5\lambda-2\beta-1\delta$	0.996226*	-208	-192	-171

each individual participant's data. The averaged estimates based on each participant's individual data are included in Appendix 2. Paired-t tests showed that two β values derived through the $4\lambda-2\beta-1\delta$ model are significantly different ($\beta_1 = 2.22 \pm 0.34 \text{ s}^{-1}$, $\beta_2 = 1.26 \pm 0.13 \text{ s}^{-1}$; $t(14) = 3.9$, $p < .01$). As for the three β values derived from the $4\lambda-3\beta-1\delta$ model ($\beta_1 = 2.21 \pm 0.34 \text{ s}^{-1}$; $\beta_2 = 1.25 \pm 0.13 \text{ s}^{-1}$; $\beta_3 = 1.30 \pm 0.13 \text{ s}^{-1}$), 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 are 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 $4\lambda-3\beta-1\delta$ model does not reflect a reliable generalization over individual subject's data, and may reflect overfitting of the empirical data. Hence we chose the $4\lambda-2\beta-1\delta$ model as the final best fit. The estimated values for each parameter in the $4\lambda-2\beta-1\delta$ model and their corresponding contrast are shown in Table 2.

To visualize the effects more clearly, we plot the data and overlay it with the function given by the $4\lambda-2\beta-1\delta$ model in two separate figures below. In Figure 2a 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 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.

As the best-fit model on the averaged data, $4\lambda-2\beta-1\delta$ 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

Table 2 Parameter estimates for the final $4\lambda-2\beta-1\delta$ model, based on parameters estimated over the average lag-latency function. 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 <i>wh</i> -Q	3.97 (3.77)	0.97	0.8
Long <i>wh</i> -Q	3.54 (3.34)	0.97	0.8

Appendix 2 gives the average parameters estimated over individual participant functions.

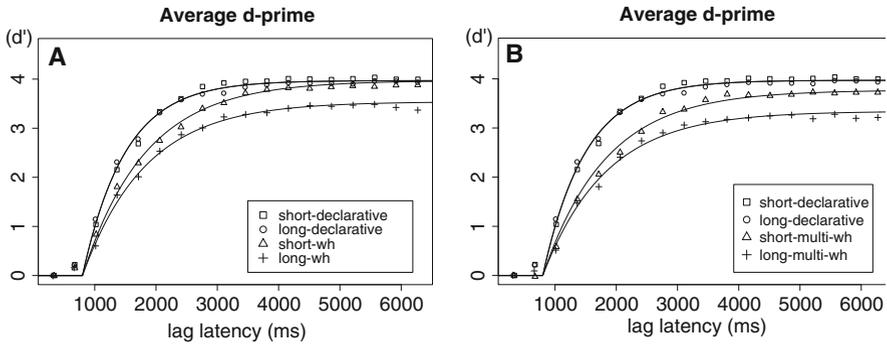


Fig. 2 a The two declarative contrasts and the two *wh*-contrasts with the semantic anomaly scaling. b The two declarative contrasts and the two *wh*-contrasts with the multiple-*wh* scaling

($p < .05$); 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 paired comparison between the two *wh*-contrasts here showed a significant difference $t(14) = 3.8$, $p < .01$.

For the rate parameter, the $4\lambda-2\beta-1\delta$ 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 ($\delta + 1/\beta$), a compound parameter referred to as the SAT function's 'speed' (Carrasco et al. 2006). This figure gives the absolute time at which a characteristic percentage of asymptotic accuracy is achieved ($\sim 63\%$; 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. It was the ultimate attained accuracy, i.e., the asymptote, that longer dependencies affect.

4 Discussion

In this experiment we compared the processing dynamics of Chinese interrogative sentences containing a *wh*-in-situ phrase with declarative sentences, both of different lengths. We did this using the SAT technique in order to decouple 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.

First, we observed a significant difference in the processing speed of declarative contrasts in comparison to interrogative contrasts: Interrogatives were processed more slowly than declaratives. *Second*, declarative and interrogative contrasts were differentially affected by length manipulation. Biclausal interrogatives attained a lower asymptotic accuracy than monoclausal interrogatives, even when the source of the anomaly contrast was an entirely string- and clause-local verb-theme pairing. In contrast, declarative sentences were unaffected by 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*-phrases 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 and 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 processing 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. It has been shown that the longer comprehenders maintain a commitment to an incorrect parse, the more difficult it is to abandon that parse upon reaching a disambiguating cue (Ferreira and Henderson

1991; Tabor and Hutchins 2004). In light of these findings, it is possible that the selective length-sensitivity observed for *wh*-contrasts might reflect such a ‘digging-in’ effect: In the long-distance *wh*-question conditions, the disambiguating cue is delayed relative to the short-distance *wh*-questions. 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 early *wh*-operator cues comprehenders to pursue an interrogative interpretation of the input, and thus no garden path arises in this comparison. For this reason, we believe the present results are unlikely to reflect a difficult reanalysis from declarative to interrogative force that is selectively exaggerated in the long-distance *wh*-question configurations.

A related possibility is that comprehenders maintain expectations about the likelihood of an upcoming *wh*-element, 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 et al. (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*-phrase 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 matrix questions. However, unlike 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*-elements. In light of this result, it seems unlikely that the present SAT results are significantly driven by comprehenders’ expectations about the likelihood of upcoming *wh*-elements.

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.

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 the 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 argued 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 the fact that this is a syntactic position without any overt content.

Acknowledgments We would like to thank Brian McElree for his advice on current work. We are grateful to Mehmeti Abdullah, Peiyao Chen and Yu Kai for their assistance in data collection. The project was partly supported by the National Natural Science Foundation of China (31170970, grant to T.-M. Guo)

Appendix 1

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

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

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

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

where 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 parentheses.

	4 λ -2 β -1 δ			4 λ -3 β -1 δ		
	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)	$\lambda_1 = 3.98$ (0.06)	$\beta_1 = 2.22$ (0.34)	$\delta = 0.93$ (0.08)	$\lambda_1 = 3.99$ (0.06)	$\beta_1 = 2.21$ (0.34)	$\delta = 0.93$ (0.08)
Long declaratives (c vs. d)						
Short wh-Qs (e vs. f)		$\beta_2 = 1.26$ (0.13)			$\beta_2 = 1.25$ (0.13)	
Short multi-Wh-Qs (e vs. i)	$\lambda_2 = 3.53$ (0.17)			$\lambda_2 = 3.53$ (0.17)		
Long wh-Qs (g vs. h)	$\lambda_3 = 3.77$ (0.07)			$\lambda_3 = 3.78$ (0.08)	$\beta_3 = 1.30$ (0.13)	
Long multi- wh-Qs (g vs. j)	$\lambda_4 = 3.32$ (0.17)			$\lambda_4 = 3.32$ (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)

References

- Anderson, M., and J. Neely. 1996. *Interference and inhibition in memory retrieval*. *Memory*, 237–313. New York: Academic Press.
- Aoshima, S., C. Phillips, and A. Weinberg. 2004. Processing filler-gap dependencies in a head-final language. *Journal of Memory and Language* 51(1): 23–54.
- Aoun, J., and A.Y.-H. Li. 1993. Wh-elements in situ: Syntax or LF? *Linguistic Inquiry* 24: 199–238.
- Carrasco, M., A.M. Giordano, and B. McElree. 2006. Attention speeds processing across eccentricity: Feature and conjunction searches. *Vision Research* 46: 2028–2040.
- Chen, Z., and J.T. Hale. 2010. Deforesting logical form. In *The mathematics of language. Lecture notes in artificial intelligence*, vol. 6419, ed. C. Ebert, G. Jäger, and J. Michaelis, 13–28. Heidelberg: Springer.
- Cheng, L.L.S. 2003. Wh-in-situ. *Glott International* 7(4): 103–109.

- Dosher, B.A. 1976. The retrieval of sentences from memory: A speed–accuracy study. *Cognitive Psychology* 8(3): 291–310.
- Dosher, B.A. 1979. Empirical approaches to information processing: Speed–accuracy tradeoff functions or reaction time. *Acta Psychologica* 43: 347–359.
- Ferreira, F., and J.M. Henderson. 1991. Recovery from misanalyses of garden-path sentences. *Journal of Memory and Language* 30: 725–745.
- Fodor, J.D. 1995. Comprehending sentence structures. In *An invitation to cognitive science: Language*, vol. 1, ed. L.R. Gleitman, and M. Liberman, 209–246. Cambridge: MIT Press.
- Foraker, S., and B. McElree. 2011. Comprehension of linguistic dependencies: Speed–accuracy tradeoff evidence for direct-access retrieval from memory. *Language and Linguistics Compass* 5(11): 764–783.
- Frazier, L. 1999. *On sentence interpretation*. Dordrecht: Kluwer.
- Frazier, L., and C. Clifton Jr. 2000. On bound variable interpretations: The LF-only hypothesis. *Journal of Psycholinguistic Research* 29: 125–140.
- Frazier, L., and G.B. Flores d’Arcais. 1989. Filler-driven parsing: A study of gap filling in Dutch. *Journal of Memory and Language* 28: 331–344.
- 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 *Image, language, brain*, ed. Y. Miyashita, P. Marantx, and W. O’Neil, 95–112. Cambridge: MIT Press.
- Huang, C.-T.J. 1982. Logical relations in Chinese and the theory of grammar. Ph.D. dissertation, MIT.
- Kazanina, N., E. Lau, M. Lieberman, M. Yoshida, and C. Phillips. 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: Springer.
- Kohonen, T. 1989. *Self-organization and associative memory*. New York: Springer.
- Lewis, R.L., and S. Vasishth. 2005. An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science* 29: 375–419.
- Lewis, R.L., S. Vasishth, and J.A. Van Dyke. 2006. Computational principles of working memory in sentence comprehension. *Trends in Cognitive Science* 10(10): 447–454.
- Lidz, J., and J. Musolino. 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. In Proceedings of the international symposium on tonal aspects languages: With emphasis on tone languages, Beijing, 125–129.
- Liu, C.-C., and P.-L. Smith. 2009. Comparing time–accuracy curves: Beyond goodness-of-fit measures. *Psychonomic Bulletin and Reviews* 16(1): 190–203.
- Liu, F., and Y. Xu. 2005. Parallel encoding of focus and interrogative meaning in Mandarin intonation. *Phonetica* 62: 70–87.
- Macmillan, N.A., and C.D. Creelman. 2004. *Detection theory: A user’s guide*. New York: Cambridge University Press.
- Martin, A., and B. McElree. 2008. A content-addressable pointer mechanism underlies comprehension of verb–phrase ellipsis. *Journal of Memory and Language* 58: 879–906.
- McElree, B. 2000. Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research* 29: 111–123.
- McElree, B. 2001. Working memory and focal attention. *Journal of Experimental Psychology: Learning, Memory and Cognition* 27: 817–835.
- McElree, B., and B.A. Dosher. 1993. Serial retrieval processes in the recovery of order information. *Journal of Experimental Psychology: General* 122: 291–315.
- McElree, B., S. Foraker, and L. Dyer. 2003. Memory structures that subservise sentence comprehension. *Journal of Memory and Language* 48: 67–91.
- Meyer, D.E., D.E. Irwin, A.M. Osman, and J. Kounois. 1988. The dynamics of cognition and action: Mental processes inferred from speed–accuracy decomposition. *Psychological Review* 95: 183–237.
- Miyamoto, E.T., and S. Takahashi. 2000. The processing of *wh*-phrases and interrogative complementizers in Japanese. In *Japanese/Korean linguistics*, vol. 10, ed. N. Akatuka, and S. Strauss, 62–75. Stanford: CSLI.
- Pablos, L. 2006. Pre-verbal structure building in Romance languages and Basque. Ph.D. dissertation, University of Maryland, College Park.

- 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., M. Ford, U. Frauenfelder, and J. Bresnan. 1988. On the temporal course of gap-filling and antecedent assignment during sentence comprehension. In *Language structure and processing*, ed. B. Grosz, R. Kaplan, M. Macken, and I. Sag. Stanford: CSLI.
- Tabor, W., and S. Hutchins. 2004. Evidence for self-organized sentence processing: Digging-in effects. *Journal of Experimental Psychology: Learning, Memory and Cognition* 30: 431–450.
- Traxler, M.J., and M.J. Pickering. 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.
- Tsai, W.-T.D. 2008. Left Periphery and how-why alternations. *Journal of East Asian Linguistics* 17: 83–115.
- Ueno, M., and R. Kluender. 2009. On the processing of Japanese *wh*-questions: An ERP study. *Brain Research* 1290: 63–90.
- Van Dyke, J.A., and R.L. Lewis. 2003. Distinguishing effects of structure and decay on attachment and repair: A cue-based parsing account of recovery from misanalyzed ambiguities. *Journal of Memory and Language* 49(3): 285–316.
- Wagenmakers, E.-J. 2007. A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review* 14: 779–804.
- Wagenmakers, E.-J., and S. Farrell. 2004. AIC model selection using Akaike weights. *Psychonomic Bulletin and Review* 11(1): 192–196.
- Wagers, M. 2013. Memory mechanisms for *wh*-dependency formation and their implications for islandhood. In *Experimental syntax and island effects*, ed. J. Sprouse, and N. Hornstein. Cambridge: MIT Press.
- Wagers, M., E. Lau, and C. Phillips. 2009. Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language* 61: 206–237.
- Wagers, M., and C. Phillips. 2009. Multiple dependencies and the role of the grammar in real-time comprehension. *Journal of Linguistics* 45: 395–433.
- Wanner, E., and M.P. Maratsos. 1978. An ATN approach to comprehension. In *Linguistic theory and psychological reality*, ed. M. Halle, J. Bresnan, and G.A. Miller. Cambridge: MIT Press.
- Warren, T., and E. Gibson. 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., A.T. Corbett, and B.A. Doshier. 1980. Priming and retrieval from short-term memory: A speed-accuracy tradeoff analysis. *Journal of Verbal Learning and Verbal Behavior* 19: 387–404.
- Xiang, M., F.Q. Liu, P.Y. Chen, and T.M. Guo. 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 University.