Cognitive Science 34 (2010) 94–112

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2009.01067.x

An On-Line Study of Japanese Nesting Complexity

Kentaro Nakatani,[a] Edward Gibson[b]

[a]*Department of English and American Literature and Language, Konan University*
[b]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology*

Received 5 August 2008; received in revised form 24 June 2009; accepted 29 June 2009

Abstract

This paper reports the results of a self-paced reading experiment in Japanese in which the materials consisted of four versions of successively more nested syntactic structures. It was found that (1) people read the more nested materials slower than the less nested materials; and (2) the locus of the relative slowdown occurred early in the nested structures. There was no corresponding slowdown when processing the verbs at the end of each clause. The results are therefore not predicted by retrieval-based integration accounts of syntactic complexity. Rather, the results support expectation-based accounts of syntactic complexity for these materials.

Keywords: Sentence Processing; Nested Structures; Expectation-based parsing; Head-final language; Retrieval; Japanese; Working memory

1. Introduction

It is well known that *nested* (or *center-embedded*) structures are harder to understand than their right- or left-branching counterparts (Chomsky & Miller, 1963; Yngve, 1960). For example, the right-branching English structure in (1a) is easier to understand than the nested structure in (1b):

(1) a. Mary met the senator who attacked the reporter who ignored the president.
b. # The reporter who the senator who Mary met attacked ignored the president.

These two sentences are identical in respect to the lexical items, the propositional/semantic content, and the grammatical complexity in terms of the depth of

Correspondence should be sent to Kentaro Nakatani, Department of English and American Literature and Language, Konan University, 8-9-1 Okamoto, Higashi-Nada, Kobe 658-8501, Japan. E-mail: kentaron@konan-u.ac.jp structural embedding (i.e., they both involve doubly embedded relative clauses [RCs]); hence the difference in the processing complexity between (1a) and (1b) cannot be attributed to the differences in syntax or semantics (in a narrow sense), or factors such as event plausibility and lexical frequency. It has thus often been argued in the literature that the contrast between (1a) and (1b) reflects how the process of sentence comprehension is constrained by the available working memory resources. To date, several different hypotheses have been proposed as to how working memory resources constrain on-line sentence comprehension. According to *retrieval-based* hypotheses, processing an incoming word *w* causes the processor to retrieve relevant pieces of the current structure, so that *w* can be integrated into this structure. The difficulty of performing a retrieval operation may vary according to several factors, such as how far back the to-be-retrieved element is in the input stream and how well this element matches the cues of the retrieval site. A number of retrieval-based theories have been proposed in the literature, including the integration component of the dependency locality theory (DLT; Gibson, 1998, 2000), the cue-based retrieval theory of Van Dyke and Lewis (2003) and Van Dyke and McElree (2006), the interference hypothesis of Gordon, Hendrick, and Johnson (2001, 2004), and the time-based activation decay model of Lewis and Vasishth (2005) and Vasishth and Lewis (2006). Some of the evidence for the retrieval-based accounts has come from reading studies in which reading times (RTs) at a verb increase depending on the distance back to verb's dependent(s) (Gibson, 1998; Gordon et al., 2001; Grodner & Gibson, 2005; King & Just, 1991). In doubly nested structures similar to (1b), for example, Grodner and Gibson (2005) showed that RTs were slowest on the outer verbs in the structure, those which are the endpoints of the longest dependencies, such as *attacked* and *ignored* in (1b).

According to an alternative class of hypotheses—the *expectation-based* hypotheses—processing difficulty is incurred either when a new expectation is added to working memory (e.g., as proposed in the storage component of the DLT; Gibson, 1998, 2000), when the input does not match the comprehenders' expectations (the anticipation hypothesis by Konieczny & Döring, 2003; the probabilistic parsers of Hale, 2001; Levy, 2005, 2008), or both. We will now elaborate on the two different expectation-based accounts. According to the storage component of the DLT (Gibson, 1998, 2000; cf. Chomsky & Miller, 1963; Gibson, 1991), the human sentence processor is sensitive to the number of syntactic heads that are required to form a grammatical sentence at each processing state. Thus, for example, following the words *the reporter who the senator who Mary…* in (1b), there is a cost of five expectation units: one for each predicted verb at this point (three such predictions) and one for each noun phrase (NP) gap position associated with each RC pronoun *who* (two such predictions).[1] In (1a), the prediction storage cost does not surpass 2 units (or maybe one) at any point of the sentence, hence the processing load is much smaller than (1b). Chen, Gibson, and Wolf (2005) and Gibson, Desmet, Grodner, Watson, and Ko (2005) provide evidence for the prediction storage cost from on-line reading experiments in English. For example, Chen et al. (2005) investigated the processing of embedded English clauses with zero, one, or two further verbs pending, as in (2):

(2) a. 0 expected verbs:
The employee realized that the boss implied that *the company planned a layoff* and so he sought alternative employment.
b. 1 expected verb, late:
The employee realized that the implication that *the company planned a layoff* was not just a rumor.

c. 1 expected verb, early:
The realization that the boss implied that *the company planned a layoff* caused a panic.
d. 2 expected verbs:
The realization that the implication that *the company planned a layoff* was not just a rumor caused a panic.

The critical region in this design consists of the embedded clause *the company planned a layoff*, in italics. Because this clause has the same structure in all conditions, integration costs are identical across the four. In sentence (2a), the critical material *the company planned a layoff* is embedded as the sentential complement of the verb *implied*, which is itself part of a clause embedded as the sentential complement of the matrix verb *realized*. Because both verbs *implied* and *realized* are encountered immediately after their respective subject nouns, no additional verbs are expected after the critical embedded clause. In sentence (2b), the verb *implied* is nominalized to *implication* with the result that the critical clause is a sentential complement of the noun *implication*. This change to the embedded subject NP *the implication* results in the requirement for an additional verb following the critical region. Similarly, in sentence (2c), the matrix verb *realized* is nominalized to *realization*, leading to the expectation for an additional verb after the critical region. Finally, in sentence (2d), both the verbs *realized* and *implied* are nominalized and two verbs are therefore required following the critical region. As predicted by syntactic expectation costs, the critical region was read fastest in (2a), slower in (2b) and (2c), and slowest in (2d), with all predicted differences significant.

A second kind of expectation-based cost has been proposed in the literature: the cost for processing a less expected input. Hale (2001) and Levy (2005, 2008), for example, propose the *surprisal* hypothesis, a theory of expectations based (solely) on structural frequencies. According to the surprisal hypothesis, all the possible structures compatible with the word string encountered so far ($w_{1...i}$) are allocated memory resources according to their relative frequencies; the next input w_{i+1} may eliminate from consideration some of the structures that are compatible with $w_{1...i}$, updating the probability distribution; w_{i+1} is difficult to process if some structures that are relatively highly frequent given $w_{1...i}$ are rejected by w_{i+1} (i.e., w_{i+1} is not highly expected). Under this hypothesis, multiple center-embedded RCs like (1b) are hard because having an object-extracted RC after a subject NP is relatively infrequent, and having such a sequence twice in a row is even more infrequent.

The goal of this study is to test the predictions of the retrieval- and the expectation-based hypotheses in Japanese, a strictly head-final language, for processing nested sentences comparable to English (1a) and (1b). Whereas it is known from previous studies of Japanese sentence processing that nested structures are harder to process than less nested ones in off-line acceptability judgment tasks (Babyonyshev & Gibson, 1999; Lewis & Nakayama, 2002; Uehara, 1997; Uehara & Bradley, 2002), it is not known how nested Japanese structures compare with their nonnested controls in on-line processing. Although both retrieval- and expectation-based theories predict nested structures to be more complex, the two classes of theories make different predictions as to the locus of the complexity, as we discuss next.

1.2. Double nesting in Japanese

Because an object precedes the verb in a head-final language, a complement clause (CC, analogous to the object of a verb) in Japanese comes between the subject and the verb. By recurring this Subject–CC–Verb frame, it is possible to create multiple-nested structures.[2] In addition, Japanese allows scrambling of its arguments, so a nonnested control can be created by placing the subject immediately next to the verb, yielding a CC–Subject–Verb order. This study tested doubly nested CC structures, which represented the basic word order, against their scrambled variants with single nesting or no nesting, as in (3).

(3) a. Doubly nested: [NP1 [NP2 [NP3 V3 comp] V2 comp] V1]
[syoki-ga [daigisi-ga [syusyoo-ga utatanesita to] koogisita to] hookokusita]
[secretary-nom [congressman-nom [prime minister-nom dozed comp] protested comp] reported]
‘‘The secretary reported that the congressman had protested that the prime minister had dozed.’’

b. Singly nested version 1: [NP2 [NP3 V3 comp] V2 comp] [NP1 V1]
[daigisi-ga [syusyoo-ga utatanesita to] koogisita to] syoki-ga hookokusita]
[congressman-nom [prime minister-nom dozed comp] protested comp] secretary-nom reported]

c. Singly nested version 2: [NP1 [NP3 V3 comp] [NP2 V2 comp] V1]
[syoki-ga [syusyoo-ga utatanesita to] [daigisi-ga koogisita to] hookokusita]
[secretary-nom [prime minister-nom dozed comp] [congressman-nom protested comp] reported]

d. Non-nested: [NP3 V3 comp] [NP2 V2 comp] [NP1 V1]
[syusyoo-ga utatanesita to] [daigisi-ga koogisita to] [syoki-ga hookokusita]
[prime minister-nom dozed comp] [congressman-nom protested comp] [secretary-nom reported]

The structure in (3a) is a doubly nested, nonscrambled (default) word order for the clause ‘‘the secretary reported that the congressman protested that the prime minister dozed.’’ The other three word orders are scrambled versions of (3a), which are less nested but have the identical propositional content. In (3b), the highest level subject NP (NP1) and the embedded complex sentence ([NP2 [NP3 V3 comp] V2 comp]) are scrambled from their default positions as arguments of V1, so that the subject NP1 is now adjacent to its thematic role-assigning verb V1. In (3c), the middle subject NP (NP2) and the sentential complement ([NP3 V3 comp]) are scrambled from their default positions, so that NP2 is now adjacent to its role-assigning verb V2. In (3d), both of the preceding scramblings take place, resulting in a fully nonnested structure, in which all arguments are adjacent to their role-assigning verbs.[3,4]

An advantage of comparing nested to non-nested CC structures in Japanese over comparing nested to nonnested RC structures in English is that the clausal embedding position of the different NPs is better controlled in Japanese than in English. That is, the main clause subject NP is the same in all versions of the Japanese sentences, as are the embedded subject and the most embedded subject. In contrast, the embedding status of the verbs and NPs differs across the English structures being compared in (1) (e.g., *Mary met the senator …* is the top-level clause in (1a), whereas *The reporter … ignored the president* is the top-level clause in (1b)).

1.3. Predictions

First, consider the predictions of the retrieval-based hypothesis for (3a–d). For simplicity, consider Gibson's (1998) linear distance-based hypothesis, whereby retrieval cost increases by one unit for each nonpronominal NP and verb intervening between the retrieval site and the to-be-retrieved word (Table 1; other retrieval-based hypotheses make essentially the same predictions). Specifically, Gibson (1998) proposes a retrieval cost metric such that

Table 1
Predictions of retrieval-based accounts

Region	1	2	3	4	5	6
(3a)	[NP1	[NP2	[NP3	V3 comp]	V2 comp]	V1 conj]
New NP/verb	1	1	1	1 0	1 0	1 0
Integration distance	0	0	0	0 0	2 0	4 0
Total cost (per region)	1	1	1	0.5	1.5	2.5
(3b)	[NP2	[NP3	V3 comp]	V2 comp]	[NP1 t	V1 conj]
New NP/verb	1	1	1 0	1 0	1	1 0
Integration distance	0	0	0 0	2 0	0	0 0
Total cost (per region)	1	1	0.5	1.5	1	0.5
(3c)	[NP1	[NP3	V3 comp]	[NP2 t	V2 comp]	V1 conj]
New NP/verb	1	1	1 0	1	1 0	1 0
Integration distance	0	0	0 0	0	0 0	4 0
Total cost (per region)	1	1	0.5	1	0.5	2.5
(3d)	[NP3	V3 comp]	[NP2 t	V2 comp]	[NP1 t	V1 conj]
New NP/verb	1	1 0	1	1 0	1	1 0
Integration distance	0	0 0	0	0 0	0	0 0
Total cost (per region)	1	0.5	1	0.5	1	0.5

each new word incurs a cost of one unit if it is the head of an NP or verb phrase that corresponds to a new discourse referent, and then one additional unit for each new NP or verb between the retrieval site and the to-be-retrieved element(s) in the structure for the input. For example, at regions 1, 2, and 3 of structure (3a), a new NP is connected to the structure for the input. In each case, the NP is connected immediately following the structure for the preceding word in the input. Thus, the total cost of each of these integrations is one unit, corresponding to the new NP in each case. At region 4, V3 is connected to NP3, corresponding to another local integration of a cost of one unit. The complementizer integrates following V3, also locally (no new NPs or verbs intervene), so this connection is cost-free according to this metric. The average cost over this region is therefore 0.5 units. At region 5, V1 is connected to NP2 at a cost of 3 units: One for each of V3 and NP3, both of which come between V2 and NP2, and one for V2 itself. There is no cost for the local integration of the complementizer, and thus the average integration cost in this region is 1.5 units. The calculation of integrations costs proceeds similarly for the remainder of the positions in the table.

According to this metric, in (3a), integration costs should be largest at the outermost verb, V1, and progressively smaller on the more embedded verbs. Note that V1 is contained in the final region (region 6) in all four conditions, making the cross-condition comparisons straightforward at this region. The processing of V1 in (3c) should be as complex as V1 in (3a), because it integrates with N1, which is the sentence-initial argument, just as in (3a). Region 5 of (3a) and region 4 of (3b) are predicted to have the same retrieval costs (2 units for a verb and a noun). All other verbs are integrated locally, causing no distance-based integration costs.

Now consider the two types of expectation-based theories: storage and surprisal. According to the syntactic storage cost hypothesis of Gibson (1998, 2000), each new nominative NP is associated with the expectation of a verb to come.[5] The detailed expectation predictions for (3a–d) according to this hypothesis are provided in Table 2. In (3a), one verb is predicted at the NP1-nom region. Then NP2-nom triggers the prediction of another verb,

Table 2
Predictions of the syntactic prediction storage cost account

Region	1	2	3	4	5	6
(3a)	[NP1	[NP2	[NP3	V3 comp]	V2 comp]	V1 conj]
Predicted heads	V1	V1,V2	V1,V2,V3	V1,V2 V1,V2	V1 V1	– V
Cost (per region)	1	2	3	2	1	0.5
(3b)	[NP2	[NP3	V3 comp]	V2 comp]	[NP1	V1 conj]
Predicted heads	V2	V2,V3	V2 V2	– V1	V1	– V
Cost (per region)	1	2	1	0.5	1	0.5
(3c)	[NP1	[NP3	V3 comp]	[NP2	V2 comp]	V1 conj]
Predicted heads	V1	V1,V3	V1 V1	V1,V2	V1 V1	– V
Cost (per region)	1	2	1	2	1	0.5
(3d)	[NP3	V3 comp]	[NP2	V2 comp]	[NP1	V1 conj]
Predicted heads	V3	– V2	V2	– V1	V1	– V
Cost (per region)	1	0.5	1	0.5	1	0.5

adding to the expectation cost. Furthermore, another verb is predicted at NP3-nom, augmenting the expectation cost to 3 units. At the following regions, the storage cost decreases as the predicted heads are encountered. In the singly nested conditions (3b) and (3c), on the other hand, the expectation cost never elevates to 3 units. At the first region, one verb is predicted, and at the second region, another verb is predicted, increasing the expectation cost to 2 units; however, at the following verb region, the cost is reduced to 1 unit because one of the two predicted heads is encountered. In (3b), V2 at the following region further reduces the expectation cost to 0; the complementizer at the same region, however, triggers a new prediction for a new verb.[6] The complementizer was presented in the same region as the preceding verb. As a simplifying assumption, we quantify the expectation cost of a region containing two words (e.g., a verb and a following complementizer) as the average expectation cost across the two words. Thus, the cost for the fourth region of condition (3b) is (0 + 1)/2 = 0.5, as shown in Table 2.

Finally, let us consider the predictions of the surprisal hypothesis (Hale, 2001; Levy, 2008). The predictions of relative difficulty are generally similar to the predictions of the storage hypothesis, especially at the NP regions, although they may be somewhat different at some V regions. According to the surprisal hypothesis, each of the possible structures compatible with the partial string that has been encountered so far is allocated a certain amount of memory resources in accordance with its probability relative to the others. Because Japanese is an SOV language and because CC-taking verbs such as *omow* ''think'' are frequent, the nested structure frame in the form of S–CC–V is a frequent structure. However, a structure containing two consecutive phonetically overt nominative NPs is not so frequent, because embedded CC structures may be scrambled and/or they may include phonetically null pronouns. According to a corpus count, which we conducted on the 23,788 occurrences of non-null nominative nouns in the Kyoto Text Corpus (a tagged corpus from the newspaper Mainichi Shimbun from 1995, consisting of about 40,000 sentences), a nominative-marked noun was immediately followed by another noun (rather than a verb, adjective, etc.) in linear order 6,722 times (28.3%), and among these cases, it was immediately followed by another nominative-marked noun only 60 times = 0.9% of all nouns following a nominative noun.[7] There were no cases of three consecutive nominative nouns (0%).[8] Thus, it is probably safe to assume that the surprisal hypothesis would predict a slowdown effect when a nominative NP in (3) is immediately followed by another nominative NP, and a greater slowdown when two consecutive NPs are followed by yet another nominative NP in (3a).

As for the V regions, the surprisal hypothesis predicts that all of the V regions are easier to process than the second of the consecutive nominative NPs, because each of the Vs are preceded by at least one of their arguments, which should create an expectation for the Vs, whereas there is no structural factor triggering an expectation of a nominative NP coming immediately after a nominative NP. This prediction is similar, but not identical, to that of the storage hypothesis: In contrast, the latter predicts that one V region, namely V3 in (3a), should be as hard to process as the second of the consecutive nominative NPs, because in both cases, two predictions are not yet matched. Furthermore, the surprisal hypothesis may predict that the V3 region in (3a) has lower surprisal than the

V3 regions in the other conditions (3b-d), because it is preceded by more context (3 NPs) compared to V3 in (3b,c) (preceded by two NPs) and V3 in (3d) (preceded by one NP and thus predicted to be the hardest);[9] the storage hypothesis, on the other hand, predicts the opposite: V3 in (3d) (2 storage units) should be harder than the other V3 regions (0.5 storage unit).

In summary, for the NP regions, the retrieval-based theories do not predict differences, whereas the expectation-based theories predict increasing RTs as nominative NPs stack up. For the V regions, the retrieval-based theories predict increasing RTs in longer distance connections, whereas the expectation-based theories do not predict such increasing RTs: The storage hypothesis predicts speedup at the V regions as more verbs are encountered, irrespective of dependency length. The surprisal hypothesis predicts that all V regions, including V3 in condition (3a), are easier than the second of the consecutive NPs, whereas the storage hypothesis predicts that V3 in (3a) is as hard as the NP regions with the storage cost of 2 units. The storage hypothesis also predicts that V3 in (3a) is harder to process than the other V3 regions, whereas the surprisal hypothesis predicts that V3 in (3a) may be the easiest among all the V3 regions because it has more preceding context and that V3 in (3d) may be the hardest among them.

2. Experiment

2.1. Method

2.1.1. Participants

Forty-five adult native speakers of Japanese in the Boston area participated in the experiment. They were each paid five dollars for participation in the experiment, which took about 20 min per session.

2.1.2 Materials

A sample set of the four target conditions described previously is shown in (3) previously. Because people typically read the final region of a sentence more slowly than other regions, we did not want to compare end-of-sentence regions to nonsentence-end regions. Consequently, the materials in the experiment included an additional clause following the sentences in (3), initiated by a conjunction such as *node* ''because'' (see Appendix S1 for a list of the materials). There were six regions of interest in the experiment, corresponding to the three subject nouns and their corresponding verbs. Bound morphemes such as case-markers, complementizers, and conjunction markers were grouped with preceding words such as nouns and verbs.

2.1.3. Procedure

The experiment was conducted with Linger, a sentence processing experimental presentation program written by Douglas Rohde, using Apple PowerBook computers on Mac OS X (Cupertino, CA). The program presented one sentence at a time on the computer monitor, left to right, word by word in a noncumulative, moving-window manner as a participant pushed the space bar (Just, Carpenter, & Woolley, 1982). The 20 sets of four target conditions were distributed in a Latin Square design, resulting in four lists. Seventy-two filler items were added to each list, including 48 items from two unrelated experiments and 24 items which were unrelated to any of the experimental items in any of the subexperiments. The 92 sentences in a list were presented in a different pseudo-random order for each participant, such that no two target items were presented consecutively. The participants were asked to read the sentences as naturally as possible. The experiment was preceded by brief instructions and nine practice items. Each stimulus was immediately followed by a yes–no question (e.g., ''Was it the congressman that protested?'' for (3)) regarding the content of the sentence was presented, with feedback for wrong answers.

2.2. Results

2.2.1. Comprehension performance

Because the nature of our materials (three clauses, many doubly embedded) made them difficult to comprehend, participants achieved a relatively low overall accuracy rate of 72.2%. Broken down by conditions, the comprehension question response accuracy rates were 77.3% (SE 2.8) for (3a), 70.7% (SE 3.0) for (3b), 69.8% (SE 3.1) for (3c), and 71.1% (SE 3.0) for (3d). A repeated-measures one-way ANOVA revealed no significant differences among the four conditions (*F*s < 2, *p*s > .2). It should be noted that the accuracy on the 24 filler sentences that were not part of any subexperiment was much higher than in the target sentences, at 87.0% across the experiment. This higher accuracy rate demonstrates that the participants were making an effort to comprehend the items. The low comprehension accuracy rate of the target items probably reflects the complexity of the items rather than the negligence of the participants.

2.2.2. Reading times

In order to reduce RT differences due to region length and to diminish RT differences among participants, length-adjusted (residual) RTs (Ferreira & Clifton, 1986) were calculated per region. It should be noted that the Japanese orthography consists of two types of characters: kana and kanji. Kana are mostly moraic, with each kana character usually corresponding to a single mora (a phonological unit determining syllabic weight in Japanese); on the other hand, kanji (Chinese characters) used in the Japanese orthography may represent several morae (ranging from one to three morae in the majority of cases). Because of this property of the Japanese orthography, a simple count of the characters used may not be a good estimate of how long people take to read the character, if there is phonological component to reading. Some previous experimental studies have shown that both the number of characters and the number of morae have an effect in reading sentences in Japanese (e.g., Mazuka, Itoh, Kondo, & Brown, 2000). We thus calculated the residual RTs for each participant using a linear regression equation involving both the number of characters and the number of morae as variables (following Mazuka, Itoh, & Kondo, 2002). The analyses of the raw RTs and of the residual RTs based solely on the character count showed similar patterns as the analyses of the dual-factor residuals (based on the mora count and the character count). We present the analyses of the dual-factor residuals here.

In addition, the residual RTs were trimmed so that data points beyond 5 *SD* from the relevant condition × region cell mean were discarded, corresponding to <1.0% of the data. We also excluded the data from one participant whose comprehension accuracy rate was below 65%. We analyzed all the other trials, whether the comprehension question was answered correctly or not. The analyses without the incorrect responses showed very similar statistical patterns.

The mean residual RTs per region for the four conditions are illustrated in Fig. 1. The residual RTs and the predictions of the retrieval-based integration and the syntactic prediction storage cost hypotheses are summarized in Table 3.

Correlational analyses conducted between the predicted costs and the overall condition mean RTs in each region (obtained by simply averaging the RTs for all trials in a region, for all participants and all items) across the four conditions support the expectation-based storage cost hypothesis. We found a highly reliable correlation between the mean residual RTs and the predicted costs of expectation (r^2 = .80; p < .001; see Fig. 2). On the other hand, when the mean residual RTs were compared to the retrieval-based integration predictions, no correlation was found (r^2 = .01; p > .6). Because the retrieval-based integration hypothesis does not make predictions for the noun regions, we also analyzed the verb regions alone, but again no correlation was observed (r^2 = .05; p > .4; see Fig. 3). In fact, the numerical trend of the direction of the correlation was in the reverse of the predicted direction.

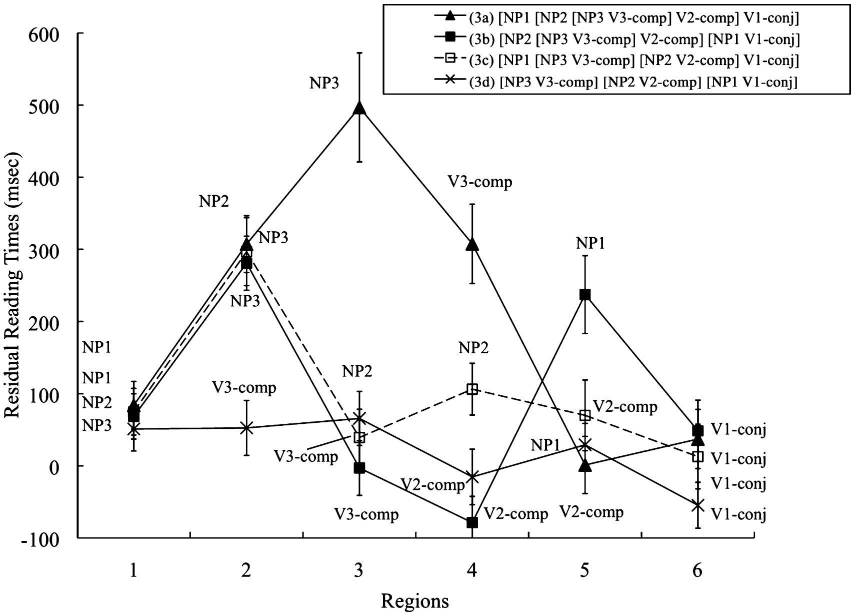

Fig. 1. Mean residual reading times for all conditions.

Table 3
Mean residual reading times in milliseconds and predicted costs

(3a)	[NP1	[NP2	[NP3	V3 comp]	V2 comp]	V1 conj]
RTs (SE)	83.9 (33.1)	307.3 (39.5)	496.8 (75.6)	307.7 (55.0)	1.2 (39.6)	37.2 (41.0)
Retrieval-based integration cost	1	1	1	0.5	1.5	2.5
Prediction storage cost	1	2	3	2	1	0.5
(3b)	[NP2	[NP3	V3 comp]	V2 comp]	[NP1 t	V1 conj]
RTs (SE)	68.5 (31.4)	280.8 (37.5)	−3.0 (38.0)	−78.7 (36.3)	237.4 (54.0)	48.5 (42.4)
Retrieval-based integration cost	1	1	0.5	1.5	1	0.5
Prediction storage cost	1	2	1	0.5	1	0.5
(3c)	[NP1	[NP3	V3 comp]	[NP2 t	V2 comp]	V1 conj]
RTs (SE)	75.0 (32.3)	296.9 (47.1)	39.3 (39.2)	106.3 (35.8)	70.0 (49.0)	12.8 (44.8)
Retrieval-based integration cost	1	1	0.5	1	0.5	2.5
Prediction storage cost	1	2	1	2	1	0.5
(3d)	[NP3	V3 comp]	[NP2 t	V2 comp]	[NP1 t	V1 conj]
RTs (SE)	51.1 (30.4)	52.6 (38.0)	65.7 (37.4)	−15.3 (38.5)	29.1 (29.7)	−54.7 (31.9)
Retrieval-based integration cost	1	0.51	1	0.5	1	0.5
Prediction storage cost	1	0.5	1	0.5	1	0.5

Having found a significant correlation between prediction storage costs and the RTs, we also performed planned comparisons of all pairs of prediction storage cost regions that differed minimally, that is 0.5 versus 1 unit of prediction storage cost; 1 versus 2 units of prediction storage cost; and 2 versus 3 units. ANOVAs confirmed that there were statistically significant differences depending on the prediction storage cost factor: 0.5 versus 1 unit: $F_1(1,43) = 9.09$, $p < .005$; $F_2(1,19) = 8.19$, $p < .02$;[10] 1 versus 2 units: $F_1(1,43) = 41.16$, $p < .001$; $F_2(1,19) = 79.43$, $p < .001$; 2 versus 3 units: $F_1(1,43) = 7.66$, $p < .01$; $F_2(1,19) = 4.22$, $p = .054$. The mean RTs are summarized in Table 4.

We also performed the same comparisons within the NP regions and within the V regions. As for the NP regions, the prediction storage cost factor yielded significant differences: 1 versus 2 units: $F_1(1,43) = 28.49$, $p < .001$; $F_2(1,19) = 39.57$, $p < .001$; 2 versus 3 units: $F_1(1,43) = 8.98$, $p < .006$; $F_2(1,19) = 4.69$, $p < .05$. As for the V regions, the difference between the regions with 0.5 unit and those with 1 unit was not significant (Fs < 1.6, ps > .2). The small prediction storage cost difference (0.5 unit) might not have been large enough to yield a difference, for the fewer trials in comparison to when NP regions were also considered. The difference between 1 and 2 units in the V regions, on the other hand, turned out to be statistically significant [$F_1(1,43) = 19.13$, $p < .001$; $F_2(1,19) = 19.96$, $p < .001$].

Finally, we performed comparisons between the V regions in order to test the following predictions of the surprisal hypothesis and the storage hypothesis: (i) the surprisal hypothesis predicts that all V regions, including V3 in (3a), are easier than the second of the two

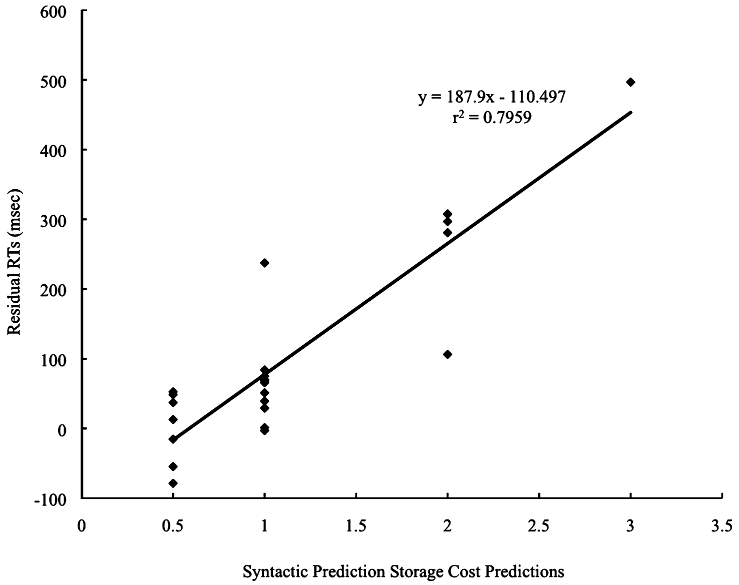

Fig. 2. Correlation between the mean reading times and the syntactic prediction storage cost predictions.

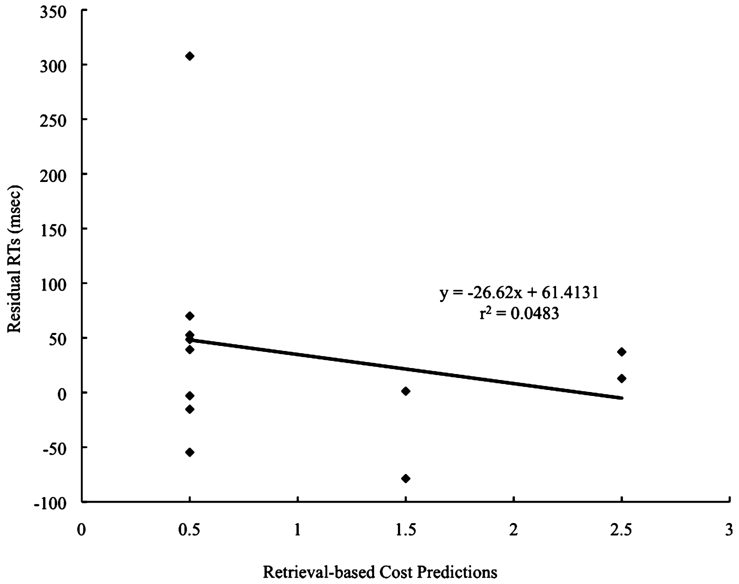

Fig. 3. Correlation between the mean reading times of the verb regions and the retrieval-based integration cost predictions.

Table 4
Predicted prediction storage costs and mean residual reading times in milliseconds

Prediction storage cost	0.5	1	2	3
Mean RT (SE)	0.4 (14.8)	65.4 (11.6)	259.7 (19.6)	496.8 (75.6)

consecutive nominative NPs, whereas the storage hypothesis predicts that V3 in (3a) is as hard as them; (ii) the surprisal hypothesis predicts that V3 in (3a) may be the easiest and V3 in (3d) may be the hardest among the V3 regions, whereas the storage hypothesis predicts that V3 in (3a) is harder than the others. As for (i), a simple comparison among all the V regions and all the regions containing the second of the two consecutive nominative NPs supported the prediction that the V regions are faster [mean RTs: 41.2 ms (SE 12.4) vs. 295.0 ms (SE 24.0); $F_1(1,43) = 32.45$, $p < .001$; $F_2(1,19) = 49.29$, $p < .001$]; however, V3 in condition (3a) was not faster than the second of the two consecutive NPs (mean RTs: 307.7 ms [SE 55.0] vs. 307.3 ms [SE 39.5]; *F*s < .1, *p*s > .9), which was not predicted by the surprisal hypothesis and was predicted by the storage hypothesis. As for (ii), the comparison between the V3 regions with regard to the number of preceding words (V3 in (3a) preceded by 3 NPs; V3 in (3b,c) preceded by 2 NPs; V3 in (3d) preceded by 1 NP) revealed a significant difference [$F_1(2,86) = 12.97$, $p < .001$; $F_2(2,38) = 10.98$, $p < .001$], though in the opposite direction of the prediction, with V3 in (3a) being much slower than the others (mean RTs: (3a) 307.7 ms vs. (3b,c) 18.1 ms vs. (3d) 52.6 ms). This difficulty of V3 in (3a) is predicted by the storage hypothesis, because its predicted storage cost is 2 units while the other V3 regions are associated with 0.5 storage unit.

3. Discussion

Overall, the results support the predictions of the expectation-based theories: The participants slowed down as nominative NPs stacked and read faster at the verb regions, as both the storage and the surprisal hypotheses predicted. The correlational analyses also strongly supported the expectation cost hypothesis. Before proceeding to discuss the results in more detail, it is worth pointing out that there may be some cost associated with processed scrambled word orders in a null context, such as in our singly nested and nonnested conditions. However, even if scrambling does adversely affect RTs, it appears that expectation costs play a larger role in the processing of these materials, because the default (doubly nested) word order condition is slowest to process.

As for how well the two expectation-based hypotheses predicted RTs at the V regions, the results were mixed. The storage predictions were supported when all the regions were compared, when the NP regions were compared, and when the V regions with the storage cost of 2 and the V regions with the cost of 1 were compared, except that there was only a numerical (nonsignificant) difference between the V regions with the storage cost of 1 and those with the storage cost of 0.5. Turning to the predictions of the surprisal hypothesis, the prediction that each V region should be easier than the second of the consecutive NPs was not supported because the V3 region in condition (3a) did not conform to this prediction;

this finding, however, was predicted by the storage hypothesis. Also, among all the V3 regions, the surprisal hypothesis predicts that V3 in (3a) may be the easiest, whereas the storage hypothesis predicts that it should be the hardest; the results from our study supported the latter prediction. Thus, overall, the comparisons between the V regions appear to be more supportive of the storage hypothesis than the surprisal hypothesis. However, a caveat is in order here: Although the observed difficulty at V3 in condition (3a) was crucial in favoring the storage hypothesis over the surprisal hypothesis, the slowness of this particular V3 might have been due to a spillover from the previous region, which was the slowest of all the target regions. Furthermore, it should be noted that the storage effects and the surprisal effects are not mutually exclusive from each other: It could be the case that both affect processing.

In contrast, the results did not support the retrieval-based theories at all: We found no evidence that the distance between a verb and its arguments affected the complexity of processing the verb. Note that the absence of the retrieval-based effects has often been pointed out in the previous studies of head-final languages: Nakatani and Gibson (2008) in Japanese; Konieczny (2000) and Konieczny and Döring (2003) in German; Vasishth (2003) and Vasishth and Lewis (2006) in Hindi. The results from this study are particularly robust in pointing to the absence of locality effects, because the semantics, the lexical choice, and the embedding order of the propositions were kept constant across all the conditions. Although the lack of retrieval-based effects is technically a null result, the fact that we could not find a main effect of locality in such well-controlled materials should be taken seriously, especially in the presence of large differences predicted by expectation-based accounts.

We will now discuss the implications of the current results for theories of sentence complexity more generally. In particular, how do we reconcile the current results with the abundant evidence of retrieval effects in other structures and/or languages? Levy (2008), for example, has suggested that all kinds of processing difficulty effects may be interpreted as surprisal: There is no cost incurred by prediction storage, or by retrieval. The nesting difficulty, according to Levy, is then caused by the infrequency of the nested structures.

However, as Levy (2008) himself notes, there are instances of attested locality effects that cannot be interpreted as surprisal effects. For example, Grodner and Gibson (2005) found that the verb *supervised* in (4a–c) was read slower when more words intervened between the subject *the administrator* and the verb:

(4) a. The administrator who the nurse *supervised* ...
b. The administrator who the nurse from the clinic *supervised* ...
c. The administrator who the nurse who was from the clinic *supervised* ...

The locality effect found here runs contrary to the surprisal prediction because the surprisal hypothesis predicts that the more words precede the verb, the more the processor expects the verb: The number of all the possible structures compatible with the partial string processed so far generally decreases as more words have been encountered, which in turn leads to the increase of the memory allocation for the target structure (the verb continuation). See also Levy, Fedorenko, and Gibson (2007), who presented two experiments, directly comparing the predictions of surprisal and retrieval-based accounts in Russian RC structures, and demonstrated that the results are best explained by retrieval difficulty.

If the retrieval locality effects are found in English and Russian, why are they not found in Japanese and other head-final languages? One possibility is that the processing strategy is parameterized depending on language-dependent parameter settings, such as the order of a head and its arguments. For example, it might be the case that the processor is heavily expectation-oriented in head-final languages so that there is no cost associated with long-distance retrieval at head positions, while it may be more retrieval-based in head-initial languages. However, evidence against this hypothesis is provided by Jaeger, Fedorenko, Hofmeister, and Gibson (2008), who found evidence for anti-locality effects even in English, for subject–verb relationships.

Another possibility is that the observed difference regarding the presence of locality effects is dependent on the type of retrieval, rather than the type of language, as suggested in Grodner and Gibson (2005) and Levy (2008). In particular, the locality effects discussed earlier were found for filler–gap relations (between a relative pronoun such as *who* and its gap position) in RCs. On the other hand, the lack of locality effects ⁄ or the presence of surprisal effects reported in the studies of head-final languages, including ours, were observed in the constructions involving dependencies between predicates and arguments ⁄ adjuncts. It is therefore possible that the retrieval locality applies only in filler–gap and other ‘‘noncanonical’’ head-dependency relations. Future research on materials with filler–gap dependencies in head-final languages will help to evaluate this hypothesis.

Notes

1. The storage component in the original formulation of the DLT includes empty categories corresponding to wh-movement. The assumption of such empty elements is not crucial in order to explain the storage cost findings that are discussed here with regard to previous work in English or the current experiment in Japanese. Future work is needed in order to determine whether empty elements associated with wh-movement are associated with processing cost independent of other stored predictions (e.g., verbs).
2. It is also possible to create multiple-nested structures in Japanese by attaching an RC to an object NP. However, such structures involve a lot of temporary ambiguity, which would lead to additional complexity associated with resolving the ambiguity.
3. A reviewer has observed that there is a fifth scrambled version of this structure, as in (i), in which the most deeply embedded clause is fronted to the beginning of the sentence:

(i) [syusyoo-ga utatanesita to] [syoki-ga [daigisi-ga *t* koogisita1to] hookokusita]
[primeminister-nom dozed comp] [secretary-nom [congressman-nom *t* protested comp] reported]

The dependencies in (i) are unlike those in (3a–d) in that some are crossed, as opposed to nested. In particular, the predicate-argument dependency headed by *koogisita* (‘‘protested’’) is interrupted by *syoki-ga* (‘‘secretary-nom’’), an argument of the higher verb *hookokusita* (‘‘reported’’). Because there is evidence that processing crossed dependencies causes processing difficulty independent of other language processing factors (Gibson & Breen, 2004; Gibson, Fedorenko, & Breen, 2006; Levy, Gibson, & Fedorenko, 2008), we limited our study here to the four-nested structures in (3).

4. An additional factor that has been shown to affect the complexity of nested structures in Japanese is whether the first subject is marked as the topic, using the case-marked ‘‘wa.’’ In particular, several previous studies have shown that the difficulty of having a sequence of multiple nominative-marked NPs in Japanese can be reduced by marking the first subject NP with ‘‘wa’’ (Miyamoto, 2002; Uehara, 1997; Uehara & Bradley, 2002). One possible explanation for this effect is that processing wa-marked NPs may simply be generally easier than processing ga-marked NPs (possibly because the frequency of wa-marked NPs). A second possibility is that the sequence NP-wa NP-ga may be more frequent than the sequence NP-ga NP-ga, and hence easier to comprehend. A third possibility is that there may be less similarity-based interference in comprehending the sequence NP-wa NP-ga than the sequence NP-ga NP-ga, independent of the production frequencies (Uehara & Bradley, 2002). Because we included no ‘‘wa’’-marked NPs in our materials, we do not address these issues here.
5. Because some Japanese predicates (mostly stative ones) allow nominative objects in addition to subjects, it is possible that only one verb will follow the second nominative NP. However, the case marker ‘‘ga’’ marks a nominative subject much more often than it marks a nominative object, perhaps especially for a human NPs. Because it is known that lexical frequencies, when heavily biased, can outweigh syntactic expectations (Gibson, 1998, 2006; MacDonald, Pearlmutter, & Seidenberg, 1994), we assume that each nominative human NP is preferentially analyzed as a subject, and thus induces an expectation for a new verb. In support of this hypothesis, see the study of Miyamoto (2002), which demonstrates that a second Japanese nominative NP tends to be interpreted as a clause boundary during on-line processing.
6. Processing the complementizer after V2 in (3b) also causes the expectation for a subject NP, but subjects are often null for S-complement verbs in natural discourse, so a null subject may also be posited here, and no storage cost should be incurred.
7. This figure includes only instances in which the *word* immediately following the nominative noun is a nominative noun, and it excludes the instances in which the *phrase* following and hierarchically closest to the nominative noun is a nominative NP premodified by a genitive NP, an adjective, a demonstrative, etc., in which case the word string would be something like N-nom N-gen N-nom, N-nom Adj N-nom, etc. If such instances were included, the figure would be greater by approximately 40. There were also about 100 instances in which two nominative NPs were separated by an adverb or a dative NP.

8. No instances of three consecutive nominative phrases or nominative words were encountered. There were only two instances in the corpus of three nominative phrases before their predicates, but these had other phrases intervening.
9. We thank a reviewer for bringing up this point.
10. The RT of region 5 in condition (2b), which was predicted to have a storage cost of 1 unit, was unexpectedly slow. However, even without this slow region, there was a statistically significant difference between 0.5 and 1 units of storage in the participants analysis [$F_1(1,43) = 5.03$, $p < .03$] and a strong tendency toward a difference in the items analysis [$F_2(1,19) = 4.16$, $p = .055$].

Acknowledgments

We would like to thank the following people for helpful discussions of and comments on this work: the audience at the 2003 CUNY Sentence Processing Conference at MIT, Ev Fedorenko, Timmy Desmet, Dan Grodner, Tadahisa Kondo, Roger Levy, Reiko Mazuka, Edson Miyamoto, Hajime Ono, Shravan Vasishth, Tessa Warren, D-Dog Watson, and two anonymous reviewers.

References

Babyonyshev, M., & Gibson, E. (1999). The complexity of nested structures in Japanese. *Language*, *75*, 423–450.

Chen, E., Gibson, E., & Wolf, F. (2005). Online syntactic storage costs in sentence comprehension. *Journal of Memory and Language*, *52*, 144–169.

Chomsky, N., & Miller, G. A. (1963). Introduction to the formal analysis of natural languages. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*, Vol. 2 (pp. 269–321). New York: Wiley.

Ferreira, F., & Clifton, C., Jr. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*, 348–368.

Gibson, E. (1991). *A computational theory of human linguistic processing: Memory limitations and processing breakdown*. Unpublished doctoral dissertation. Pittsburgh, PA: Carnegie Mellon University.

Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, *68*, 1–76.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz, & W. O'Neil (Eds.), *Image, language, brain* (pp. 95–126). Cambridge, MA: MIT Press.

Gibson, E. (2006). The interaction of top-down and bottom-up statistics in the resolution of syntactic category ambiguity. *Journal of Memory and Language*, *54*, 363–388.

Gibson, E., & Breen, M. (2004). Processing crossed dependencies in English. Poster presented at the 17th Annual CUNY Conference on Human Sentence Processing, University of Maryland, March, 25–27.

Gibson, E., Desmet, T., Grodner, D., Watson, D., & Ko, K. (2005). Reading relative clauses in English. *Cognitive Linguistics*, *16*, 313–354.

Gibson, E., Fedorenko, E., & Breen, M. (2006). Processing extraposed structures in English: Grammatical and processing factors. Poster presented at the 19th Annual CUNY Conference on Human Sentence Processing, CUNY, New York, March.

Gordon, P. C., Hendrick, R., & Johnson, M. (2001). Effects of noun phrase type on sentence complexity. *Journal of Memory and Language*, *51*, 97–114.

Gordon, P. C., Hendrick, R., & Johnson, M. (2004). Memory interference during language processing. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *27*, 1411–1423.

Grodner, D., & Gibson, E. (2005). Consequences of the serial nature of linguistic input. *Cognitive Science*, *29*, 261–291.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of North American Association for Computational Linguistics* Pittsburgh, PA: Association for Computational Linguistics.

Jaeger, T. F., Fedorenko, E., Hofmeister, P., & Gibson, E. (2008). Expectation-based syntactic processing: Anti-locality outside of head-final languages. Paper presented at the 21sh CUNY Conference on Human Sentence ProcessingUniversity of North Carolina, Chapel Hill, NC, March, 13–15.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processing in reading comprehension. *Journal of Experimental Psychology: General*, *111*, 228–238.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language*, *30*, 580–602.

Konieczny, L. (2000). Locality and parsing complexity. *Journal of Psycholinguistic Research*, *29*, 627–645.

Konieczny, L., & Döring, P. (2003). Anticipation of clause-final heads: Evidence from eye-tracking and SRNs. In P. P. Slezak (Ed.), *Proceedings of the 4th international conference on cognitive science* (pp. 330–335). Sydney: University of New South Wales.

Levy, R. (2005). *Probabilistic models of word order and syntactic discontinuity*. PhD thesis. Stanford, CA: Stanford University.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*, 1126–1177.

Levy, R., Fedorenko, E., & Gibson, E. (2007). The syntactic complexity of Russian relative clauses. Paper presented at the 20th CUNY Conference on Human Sentence Processing, UC, San Diego, March, 29–31.

Levy, R., Gibson, E., & Fedorenko, E. (2008). Expectation-based processing of extraposed structures in English, Poster presented at the 21st CUNY conference on sentence processing, University of North Carolina, March, 13–15.

Lewis, R., & Nakayama, M. (2002). Syntactic and positional similarity effects in the processing of Japanese embeddings. In M. Nakayama (Ed.), *Sentence processing in East Asian languages* (pp. 85–110). Stanford, CA: CSLI Publications.

Lewis, R., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 1–45.

MacDonald, M. C., Pearlmutter, N., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*, 676–703.

Mazuka, R., Itoh, K., & Kondo, T. (2002). Costs of scrambling in Japanese sentence processing. In M. Nakayama (Ed.), *Sentence processing in East Asian languages* (pp. 131–166). Stanford, CA: CSLI Publications.

Mazuka, R., Itoh, K., Kondo, T., & Brown, J. S. (2000). Phonological processing in reading Japanese sentences silently. Presented at the 13th Annual CUNY Conference on Human Sentence Processing, University of California, San Diego, March 30–April 1.

Miyamoto, E. T. (2002). Case markers as clause boundary inducers in Japanese. *Journal of Psycholinguistic Research*, *31*, 307–347.

Nakatani, K., & Gibson, E. (2008). Distinguishing theories of syntactic expectation cost in sentence comprehension: Evidence from Japanese. *Linguistics*, *46*, 63–87.

Uehara, K. (1997). Judgments of processing load: The effect of NP-*ga* sequences. *Journal of Psycholinguistic Research*, *26*, 255–263.

Uehara, K., & Bradley, D. (2002). Center-embedding problem and the contribution of nominative case repetition. In M. Nakayama (Ed.), *Sentence processing in East Asian languages* (pp. 257–287). Stanford, CA: CSLI Publications.

Van Dyke, J., & Lewis, R. L. (2003). Distinguishing effects of structure and decay on attachment and repair: A retrieval interference theory of recovery from misanalyzed ambiguities. *Journal of Memory and Language*, *49*, 285–413.

Van Dyke, J., & McElree, B. (2006). Retrieval interference in sentence comprehension. *Journal of Memory and Language*, *55*, 157–166.

Vasishth, S. (2003). *Working memory in sentence comprehension: Processing Hindi center embeddings*. New York: Routledge.

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language*, *82*, 767–794.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society*, *104*, 444–466.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Appendix S1. The list of all target items used in the experiment for the unscrambled condition (3a).

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.