Bilingualism: Language and Cognition 9 (1), 2006, 71–88 © 2006 Cambridge University Press doi:10.1017/S1366728905002385

Incremental interpretation in second language sentence processing*

JOHN N. WILLIAMS
University of Cambridge

The degree to which native and non-native readers interpret English sentences incrementally was investigated by examining plausibility effects on reanalysis processes. Experiment 1 required participants to read sentences word by word and to make on-line plausibility judgements. The results showed that natives and non-natives immediately computed the plausibility of the preferred structural analysis, which then affected ease of reanalysis. Experiment 2 required participants to read the same sentences word by word in order to perform a memory task. The natives showed a similar pattern of results to Experiment 1, whereas for the non-natives plausibility effects were delayed. However, the non-natives still appeared to be performing immediate syntactic reanalysis. It is concluded that syntactic processing was person- and task-independent, whereas the incrementality of interpretation was more dependent on task demands for the non-natives than for the natives.

Introduction

The most salient point to come out of sentence processing research over the last few decades, and perhaps the only point on which there is widespread agreement, is that sentences are interpreted in a highly incremental fashion. Even if there has been substantial debate over the types of information and strategies that are used to guide attachment decisions, there is widespread agreement that each incoming word is immediately attached to the evolving sentence structure. The reason for the widespread acceptance of this assumption is the ease with which it is possible to obtain "garden-path" effects. Because readers interpret sentences incrementally, in a more or less word-by-word fashion, they occasionally commit to an analysis that turns out to be incompatible with later parts of the sentence. In cases of such "garden-paths" the reader must revise the initial analysis when they encounter the disambiguating information.

But what is meant by "interpret"? For Frazier (1987) interpretation refers primarily to building the syntactic phrase marker of the sentence. On the other hand, for Just and Carpenter (1987, p. 40) "The immediacy of interpretation pervades all levels of comprehension, such as encoding a word, accessing its meaning, and determining its referent and semantic and syntactic status in the sentence". Pickering and Traxler (2000, p. 239) argue that the interpretation of a sentence is "immediately integrated with relevant background knowledge and information provided by discourse context". In other words, when listening or reading, readers update their "situation model" (Kintsch, 1988) continually as each word is processed.

Pickering and Traxler's (2000) claim is based on plausibility effects in native sentence processing. For example, Traxler and Pickering (1996) compared reading times in sentences like (1) and (2).

(1) We like the book that the author wrote unceasingly and with great dedication about while waiting for a contract.

(2) We like the city that the author wrote unceasingly and with great dedication about while waiting for a contract.

In theory-neutral terms there is a "gap" after the verb *wrote* in both of these sentences. Assuming that readers follow the "Filler-Driven strategy" (Frazier and Clifton, 1989), or in a gap-free account, the principle of "immediate association" (Pickering and Barry, 1991), when they encounter the verb they should initially interpret the filler *book*, in (1), or *city*, in (2), as its direct object. But this is not just a structural hypothesis. Traxler and Pickering found that reading times were longer in the region *wrote unceasingly* in (2) than in (1) suggesting that readers immediately computed the plausibility of the filler–gap relation at the verb. Computation of plausibility entails consulting real world knowledge, in this case about the kinds of things that you can and cannot write. Hence, there is immediacy of interpretation at the level of the situation model, even for aspects of structure that ultimately have to be abandoned, since in both (1) and (2) the relevant gap turns out to be after *about* instead. In fact reading times were longer over the disambiguating region *about while* in (1) and than in (2). Traxler and Pickering (1996) assume that it is

* Thanks to Ernest Lee for helping collect the data for Experiment 1, and to the reviewers and David Green for helpful comments and advice.

Address for correspondence
University of Cambridge, Research Centre for English and Applied Linguistics, 9 West Road, Cambridge CB3 9DP, UK
E-mail: *jnw12@cam.ac.uk*

harder to reanalyse a structure that was initially thought to be plausible than one that was implausible because readers commit more strongly to plausible analyses. Thus, in sentences of this type, plausibility effects show a "cross-over" pattern – slower reading at the verb when the filler is implausible as direct object but faster reading in the disambiguating region. This pattern is diagnostic of incremental interpretation at the level of the situation model during sentence processing. Similar results have been obtained by Pickering and Traxler (Pickering and Traxler, 1998, 2003).

Turning to L2 sentence processing, there is already a good deal of evidence for incremental, and in many cases, native-like, sentence processing in non-natives (Fender, 2001). However, this research has been concerned with structural-level garden-paths rather than plausibility effects. Whilst establishing that syntactic processing is incremental, these studies do not establish that incrementality also applies at the level of the situation model.

For example, Juffs and Harrington (1996) and Juffs (2004) showed that both native English speakers and advanced Chinese learners of English show garden-path effects in sentences such as (3), where *proved* causes processing difficulty because *water* is initially interpreted as the object of *drink* rather than the subject of *proved*.

(3) After Bill drank the water proved to be poisoned.

Hoover and Dwivedi (1998) showed that sentence (5) leads to slower reading times at the verb *gouter* than sentence (4), both for native speakers of French and advanced English learners of French.

(4) Il faisait tranquillement goûter le vin avec son fromage préféré.
He had the wine be tasted quietly with his favourite cheese.

(5) Il le faisait tranquillement goûter avec son fromage préféré.
He had it be tasted quietly with his favourite cheese.

In sentence (5), the clitic *le* is initially interpreted as the direct object of the verb *faisait*. Once again there is a strong tendency to interpret a post-verbal element as an object, but note that in this case the element in question is an empty category (Pro) which is indexed with the object pronoun. The fact that the learners processed this structure in a native-like way is remarkable given that English does not have clitics or preverbal objects. But the experiment does not tell us whether computations at the level of the situation model were native-like as well.

Williams, Möbius, and Kim's (2001) study of processing long distance dependencies in L1 and L2 provided evidence for immediate computation of plausibility even amongst the non-natives. Sentences like (6) and (7) were compared.

(6) Plausible-at-V: Which machine did the mechanic fix the motorbike with two weeks ago?

(7) Implausible-at-V: Which customer did the mechanic fix the motorbike for two weeks ago?

In both sentences there is a potential gap after the verb *fix*. According to the Filler-Driven strategy, readers should hypothesise *machine* as the direct object of *fix* in (6) and *customer* as the direct object of *fix* in (7). The question is whether the plausibility of these assignments is computed, and how that affects subsequent processing. It is known that when reading structures such as these native readers show a "filled gap effect" (Stowe, 1986) – a tendency to slow down in the post-verbal region because the gap that they originally hypothesised turns out to be filled by an unexpected noun phrase, in this case *the motorbike*. The issue here is whether the filled gap effect is affected by the plausibility of the initial filler–gap assignment. The experiment employed the "stop making sense task" (Boland, Tanenhaus, Garnsey and Carlson, 1995). Participants read sentences one word at a time and were required to indicate as soon as they thought that the sentence had stopped making sense. For the natives, rates of stop making sense decisions showed the predicted cross-over pattern – more stop making sense decisions at the verb in the Implausible-at-V condition, (7), than in the Plausible-at-V condition, (6), but the opposite pattern at and following the post-verbal noun. In other words, there was a larger filled-gap effect when the initial filler–gap assignment was plausible. Reading times on trials where no stop making sense decisions were made showed no effect at the verb, but the Plausible-at-V condition was slower at and following the noun. Overall, the results suggested that the native readers were interpreting the sentences incrementally at the level of the situation model. Similar results were obtained for Chinese, Korean and German learners of English. The similarity in the results for the natives and non-natives is remarkable given that Chinese and Korean questions do not involve gaps.

However, one aspect of the Williams et al. (2001) results did suggest potential differences between natives and non-natives with respect to the rapidity with which plausibility information is utilized. In reading times only the natives showed a plausibility effect at the post-verbal determiner, the Implausible-at-V condition being slower than the Plausible-at-V condition. The filler–gap implausibility and the disambiguating syntactic cue provided by the determiner might have triggered rapid syntactic reanalysis in the Implausible-at-V condition. In contrast, for the non-native speakers, there were no plausibility effects in reading time until the post-verbal noun where the Plausible-at-V condition was slower than the Implausible-at-V condition. This may have reflected a slight delay in the utilisation of plausibility information. However, it may also have reflected what might be referred

to as "argument competition"; that is, the process of substituting the post-verbal noun for the filler as the theme in the verb's argument structure (or making the equivalent substitution in the situation model). It is not unreasonable to suppose that this substitution process is more difficult when the initial assignment was plausible. But in this case, the difficulty in processing would be a reflection of thematic/situation model level processes, rather than processes of syntactic revision. It therefore remains unclear to what extent plausibility influenced syntactic revision processes in the non-natives.

The aim of the present Experiment 1 was to further explore plausibility effects in the post-verbal region by increasing the number of words prior to the noun, as in the following.

(8) Plausible-at-V: Which machine did the mechanic fix the very noisy motorbike with two weeks ago?

(9) Implausible-at-V: Which friend did the mechanic fix the very noisy motorbike for two weeks ago?

In the majority of the items the additional words were an intensifier and an adjective, and in two items two adjectives were used. The additional material will be referred to here as the determiner–intensifier–adjective (det–int–adj) sequence. If, for non-native readers, plausibility effects reflect argument competition, then reading times in the Plausible-at-V condition (8) should only become slower than those in the Implausible-at-V condition (9) at and following the post-verbal noun *motorbike*. If there is merely a delay in utilisation of plausibility, then plausibility effects might still be obtained prior to the noun, but later than for natives.

The second aim of the present study was to address potential task effects. The Williams et al. (2001) study and the present Experiment 1 used the stop-making sense task. Participants read the sentences word by word and pressed the space bar as soon as they thought the sentence had stopped making sense. Obviously this task forces the participants to evaluate plausibility incrementally, and may also potentially alter their whole reading strategy. In particular, plausibility information might be utilised in a different way from the way it is in normal reading. Experiment 2 of the present study therefore examined plausibility effects in a task where the participants only had to read the sentences to answer comprehension/memory questions.

Experiment 1

Participants

There were 79 participants, comprising 27 native speakers of Chinese, 26 native speakers of Spanish, 8 native speakers of Italian, and 18 native speakers of English. The Spanish and Italian speakers were combined to form a group of 34 Romance speakers. All of the participants were following graduate or postgraduate courses at the University of Cambridge. The non-natives had achieved the Cambridge Certificate of Proficiency in English or an equivalent EFL qualification, which ensures that they are linguistically functional in an academic environment. The participants' English proficiency was assessed in order to facilitate comparisons across Experiments 1 and 2. Prior to the experiment the non-native participants filled out a proficiency self-assessment questionnaire based on that used in a study by Bachman and Palmer (1989). The questions probed perceived difficulty in the following areas: grammatical competence (morphology and syntax), pragmatic competence (vocabulary, cohesion, organisation), and socio-linguistic competence (register, nativeness, nonliteral language). Bachman and Palmer showed that this test has a high reliability, with alpha levels greater than 0.75 for each component. They also showed that when the questions were phrased in terms of perceived difficulty (as opposed to ability) they were particularly good at identifying different underlying traits, and showed strong correlations with an overall ability factor. Participants rated the subjective difficulty of each trait on a four-point scale, where 1 was assigned most difficult, and 4 the least difficult. The mean rating over the 7 questions was calculated for each participant so that the higher the mean score, the higher the self-rated proficiency. The mean score, standard deviation, and range for the Romance speakers was 2.86, SD = 0.39, range = 2–3.57, and for the Chinese speakers 2.99, SD = 0.45, range = 2–3.71. There was no significant difference between the score for the two groups, $t(59) = 1.03$.

Method

Materials

The critical materials comprised 16 sentences based on those used in Williams et al. (2001) except that an intensifier and an adjective were inserted between the post-verbal determiner and the noun (except in two cases where two adjectives were used). All of the sentences involved adjunct extractions of the type in (8) and (9). There were two versions of each sentence. In the Plausible-at-V version the filler was plausible as the direct object of the verb, and in the Implausible-at-V version the filler was implausible.[1] In addition the adjective was chosen so that in the Plausible-at-V condition it could potentially modify

[1] For two of the items, because the verb allowed the dative alternation the filler could be plausibly interpreted at the verb as an indirect object. These were *Which friend did the man buy* (*the really expensive radio*), and *Which girl did the man push* (*the very nice bike*). The first of these was in the Implausible-at-V condition and the second in the Plausible-at-V condition. Therefore, even if the possibility of a second internal argument facilitated reanalysis, the effect would have been the same

the *wh*-filler. For example, in *Which girl did the man push the very nice bike into late last night?*, being nice is a potential property of a girl. If an adjective like *rusty* had been used instead then this would provide a more obvious signal that another noun was going to be mentioned, which might serve as a semantic cue to reanalysis. Since the intention was to see whether reanalysis could be triggered by structural cues prior to the noun it was decided to use adjectives that could be attributed to the filler noun as well as the post-verbal object noun. The critical items are listed in the Appendix.

The 16 items were divided into two groups of eight items each. Two presentation lists were constructed. In the first, the items from one of the item groups appeared in the Plausible-at-V condition and the items from the other group appeared in the Implausible-at-V condition. These assignments were reversed to form the second presentation list. Half of the subjects in each language group received each list.

A large number of filler items were also written. There were 38 implausible sentences comprising the following: twelve implausible declaratives in which the direct object was implausible (e.g. *The rich family burned the key in their heater late last night*), two declaratives in which the adjunct was implausible, eight implausible argument extractions with a relatively short distance between filler and gap site (e.g. *Which shop did the criminal kill in the city yesterday evening?*), and eight implausible argument extractions with a relatively long distance between filler and gap site (e.g. *Which bird did Janet's older brother Frank build at the end of the road last year?*). There were also 32 sentences in which there was no implausibility: twenty declaratives, six short argument extractions, and six long argument extractions. The sentences for each presentation list were presented in three different pseudo-random orders, with the proviso that no two sentences of the same type should occur consecutively.

Procedure

Sentences were displayed one word at a time on a computer screen. All words appeared in the same position half way down the screen near to the left-hand edge, and subjects changed each word into the next by single-clicking the left mouse button with their dominant hand. They were also instructed to respond as soon as they thought that the sentence had stopped making sense by pressing the space bar with their non-dominant hand. This response also had the effect of changing the current word into the next word. After any stop making sense decision the subject continued reading the remainder of the sentence by pressing the mouse button. All times between presentation of a word on the screen and a response were measured to millisecond accuracy by the software.

in both conditions. Of course it is possible that at the verb readers consider filler roles other than direct object/theme, regardless of the verb's subcategorisation. This possibility will be considered in the Conclusion.

Results

Stop making sense decisions

The data were scored in terms of the percentage of possible responses that were made at each position by each participant in each condition. Response rates at each position were calculated relative to the number of possible responses that could have occurred at that position (only one stop making sense decision was coded for each sentence, and in the very rare cases where more than one response was made, only the earliest was counted). For example, a participant who responded at the verb in two of the eight sentences in the Plausible-at-V condition and at the noun in a further two sentences would have a response rate of 0.25 (2/8) at the verb and 0.33 (2/6) at the noun.[2] The results for each language group are displayed in Figure 1.

As explained in the Introduction, the critical region for present purposes is the immediate post-verbal det–int–adj region. For the subjects analysis Language Group and Presentation List were between-subjects factors, and Plausibility and Position were within-subjects factors. For the items analysis Language Group, Plausibility and Position were within-items factors, and item group was a between-items factor. There was no main effect of Language, $F1(2,73) = 2.78$, $p < 0.1$; $F2(2,13) = 3.25$, nor of Plausibility, F1 and F2 both < 1.0. But there was a significant interaction between Plausibility and Position, $F(2,146) = 16.17$, $p < 0.001$; $F2(2,13) = 8.11$, $p < 0.01$. At the determiner there were more responses in the Implausible-at-V condition than the Plausible-at-V condition, but this pattern was reversed over the intensifier and adjective. This interaction was present for each language group. For the natives $F1(2,32) = 4.95$, $p < 0.05$; $F2(2,13) = 4.87$, $p < 0.05$. For the Chinese $F1(2,50) = 9.91$, $p < 0.001$; $F2(2,13) = 7.19$, $p < 0.01$. For the Romance $F1(2,64) = 6.37$, $p < 0.01$; $F2(2,13) = 4.22$, $p < 0.05$.

It seems probable that the tendency for more stop making sense decisions in the Implausible-at-V condition on the determiner is a spill-over from the effect that is present on the verb. The verb effect is hardly surprising

[2] The rationale for this scoring method was that, from the participants' point of view, once a stop making sense decision had been made on a particular trial no further responses were appropriate. Thus, for each participant the calculation of the proportion of trials on which a stop making sense decision was made at a particular position had to be made relative to the number of possible responses at that position (i.e. subtracting the number of responses made earlier in the sentence from the total number of trials in that condition).

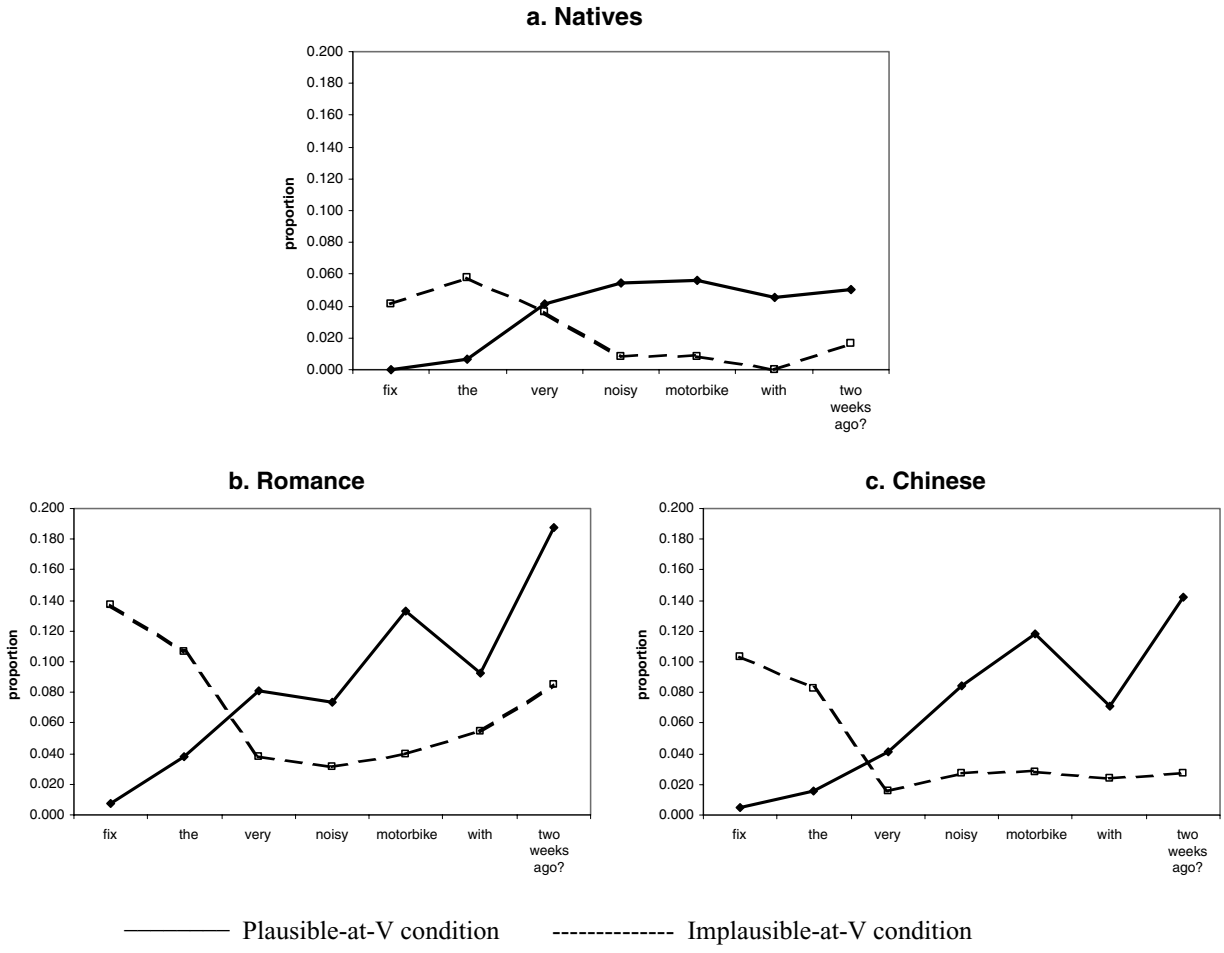

Figure 1. Experiment 1 stop making sense decisions expressed as a proportion of possible responses at each position.

given that the sentences in this condition were actually implausible at this point. Nevertheless the relatively high rate of stop making sense decisions at the verb in the Implausible-at-V condition is theoretically important because it shows that the filler was being assigned to the first gap position, triggering an immediate, or almost immediate, stop making sense decision. Analyses of variance showed that the effect of plausibility at the verb was indeed significant for all groups. For the natives, $F1(1,16) = 5.54$, $p < 0.05$; $F2(1,14) = 4.69$, $p < 0.05$. For the Chinese, $F1(1,25) = 17.09$, $p < 0.001$; $F2(1,14) = 12.12$, $p < 0.01$. For the Spanish, $F1(1,32) = 33.73$, $p < 0.001$; $F2(1,14) = 6.02$, $p < 0.05$.

Reading times

This analysis was based only on the reading times up to the point in each sentence where the participant made a stop making sense decision. Although the reading profiles over all word positions will be presented, the critical region for analysis once again comprised the det–int–adj region. Given that stop making sense decisions were made in the critical region on only a minority of trials, an analysis of reading times was deemed feasible. However, to improve the reliability of reading time estimates participants were excluded from the analysis if they made a stop making sense decision within the verb–det–int–adj–N region on more than half of the sentences in either of the conditions. The reading time analysis was based on the data from 24 Chinese, 28 Romance and 17 native English speakers.

The influence of excessively long reading times was curtailed in the following way. For each word position in the sentences a response time was classified as an outlier if it fell more than 2.5 standard deviations above the mean for all of the reading times in that position across both of the conditions. It was then replaced with the next highest value at that position, regardless of condition. This is a conservative procedure because it ignores any potential differences between conditions, whilst at the same time ensuring that the influence of long reading times is curtailed, rather than eliminated.

The general reading profiles for the three language groups were remarkably similar. Figure 2 shows the reading profiles for each sub-group of participants.

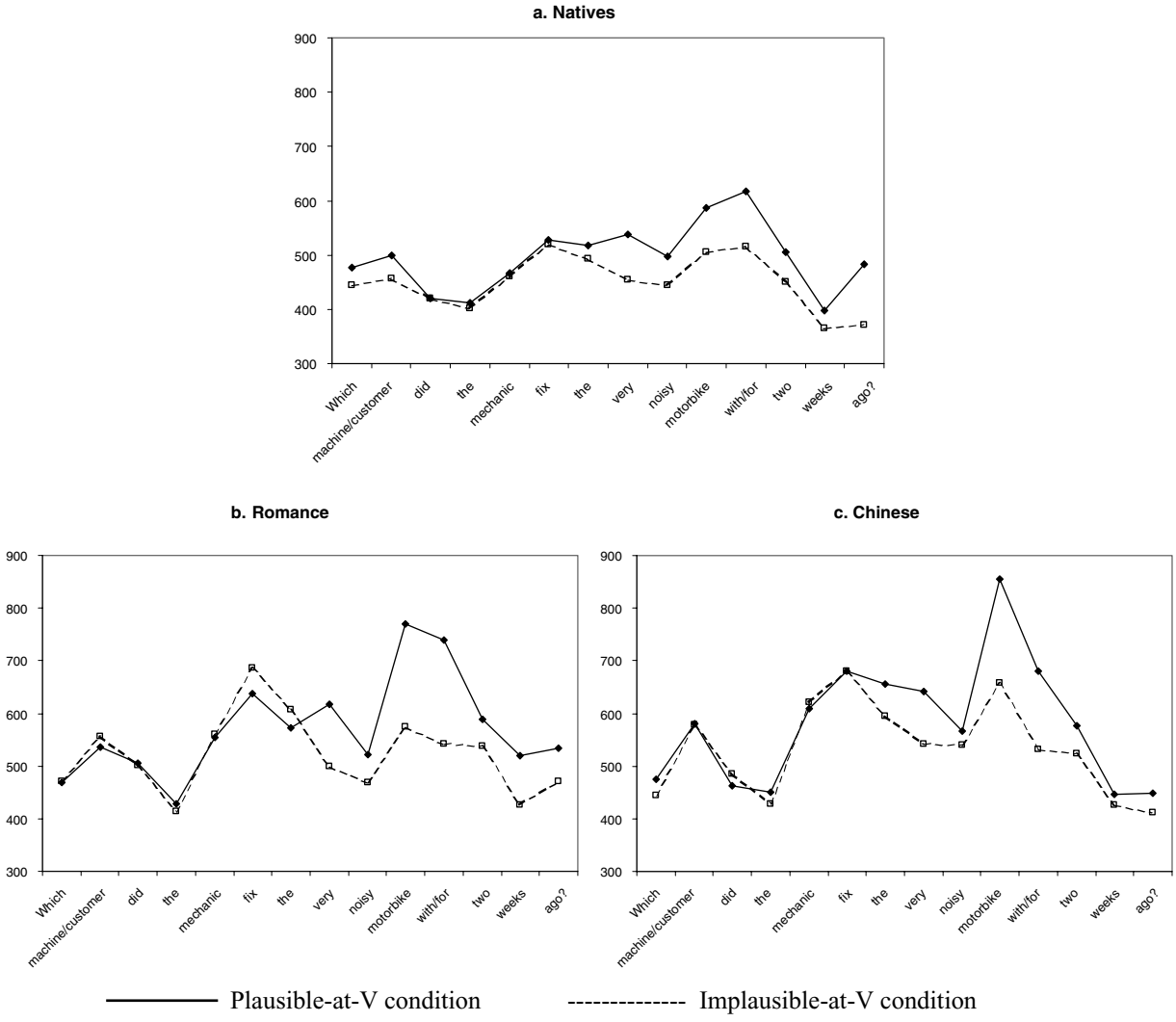

Figure 2. Experiment 1 mean reading times (in ms).

The statistical analysis was again focused on the det–int–adj region immediately after the verb. ANOVAs were performed by subjects and items.[3] There was a main effect of Plausibility, $F1(1,63) = 27.17$, $p < 0.001$; $F2(1,14) = 10.26$, $p < 0.01$, and an interaction between Plausibility and Position, $F(2,126) = 6.92$, $p < 0.001$; $F2(2,13) = 6.72$, $p < 0.01$. As shown by Figures 2a–c, reading times were slower in the Plausible-at-V condition in the det–int–adj region, with the effect being concentrated at the intensifier; hence the interaction between plausibility and position. Individual analyses on the subgroups of participants confirmed that the general effect of plausibility in the critical region was present in all language groups. For the Chinese there was a main effect of plausibility, $F1(1,22) = 10.47$, $p < 0.01$; $F2(1,14) = 6.86$, $p < 0.05$. For the Romance group there was a main effect of plausibility by subjects, $F1(1,26) = 7.07$, $p < 0.05$; but not by items, $F2(1,14) = 1.75$, and an interaction between plausibility and position, $F1(2,52) = 8.32$, $p < 0.001$; $F2(2,13) = 7.00$, $p < 0.01$. For the natives there was a main effect of plausibility, $F1(1,15) = 20.22$, $p < 0.001$; $F2(1,14) = 9.04$, $p < 0.01$. With regard to reading times at the verb there was only any hint of slower reading times in the Implausible-at-V condition for the Romance group, but even then only in the subjects analysis, $F(1,26) = 5.09$, $p < 0.05$; $F2(1,14) = 2.04$.

Discussion

There was clear evidence that all groups of participants initially interpreted the filler as the direct object of the verb in both the Plausible-at-V and Implausible-at-V

[3] For the items analysis any item which received stop making decisions by more than half of the participants in the critical region in any one condition was excluded. This resulted in the loss of one item from the native group

conditions. There was a higher rate of stop making sense decisions at and immediately following the verb in the Implausible-at-V condition (as in Boland et al., 1995). In contrast, in the Plausible-at-V condition there were more stop making sense decisions and slower reading times in the immediate post-verbal region, and crucially, prior to the post-verbal noun. This presumably reflects the increased difficulty of revising the initial structural hypothesis when that interpretation was plausible (Traxler and Pickering, 1996; Pickering and Traxler, 1998). This was the case in all participant groups. Therefore it appears that all groups were able to incrementally interpret the sentences at the level of the situation model. It was also apparent that all readers were able to interpret the det–int–adj sequence as evidence for a post-verbal noun phrase and that they used this information to trigger reanalysis. They did not have to wait until there was overt argument competition at the post-verbal noun. Both native and non-native readers therefore appeared to be equally sensitive to structural and non-structural information, and they made use of plausibility constraints with equal efficiency.

One curious aspect of the results concerns the verb in the Implausible-at-V condition. There were significantly more stop making sense decisions at this position than in the Plausible-at-V condition, as would be expected because the filler was implausible as the direct object. Yet in the reading time data there was no difference between the conditions. This means that participants were capable of explicitly responding to the implausibility that becomes apparent at the verb, but when they did not do so (which was most of the time) their reading was not disrupted. In contrast, previous research has found slower reading times at the verb when the filler–gap dependency is implausible (Pickering and Traxler, 1998; Pickering and Traxler, 2003).

The differences in the results may be due to the nature of the task. In normal reading, where readers presumably assume that all sentences will ultimately make sense, any implausibility could in theory be used as a cue to reanalyse. But here many of the sentences really did turn out to be implausible, and so it would not be surprising if readers sometimes adopted a "wait and see" strategy, not responding immediately at the verb, permitting implausible attachments to linger, and waiting for reanalysis to be forced on structural grounds. Plausibility effects would then be delayed until the post-verbal noun phrase is encountered.

Another reason for considering the nature of the task is of course that the stop making sense task forces participants to evaluate plausibility incrementally. Thus, whilst this experiment tells us that non-native readers compute the plausibility of the potential filler–gap relation at the verb, and that this influences the difficulty of reanalysis, it does not tell us that these processes operate in other reading situations that do not compel incremental interpretation at the level of the situation model.

Experiment 2 addressed the above issues by examining reading of the same critical materials in a task that simply motivated levels of comprehension that were sufficient for answering periodic questions. The same word-by-word presentation method was used as in Experiment 1, but there was no overt requirement to interpret the sentences in an incremental fashion, and all of the sentences in the experiment were plausible.

Experiment 2

Participants

A total of 33 non-native and 35 native volunteers were tested. All of the participants were following graduate or postgraduate courses at the University of Cambridge. The non-natives had achieved the Cambridge Certificate of Proficiency in English or an equivalent EFL qualification. Given that in Experiment 1 the participants' native language made no difference to the results these participants had a variety of first languages: Albanian (2), Arabic (1), Bosnian (1), Chinese (5), Croatian (4), Dutch (1), French (1), German (3), Greek (2), Korean (1), Macedonian (1), Portuguese (3), Russian (3), Singhalese (3), Spanish (1), Taiwanese (1). Their mean self-rated proficiency was 2.90, SD = 0.42, range = 2.14–3.86, which is almost identical to the proficiency of the participants in Experiment 1.

Method

Materials

Exactly the same critical materials were used as in Experiment 1, organised into two presentation lists so that each participant received eight items in each condition. The filler materials comprised 14 declarative sentences and 16 *wh*-questions (all with argument extraction). These were based on the materials from Experiment 1. Implausible fillers from Experiment 1 were altered so as to be plausible. For the comprehension/memory task a probe was written for each sentence. All probes were declarative statements with a word missing. The participants' task was to supply the missing word. For half of the critical items in each condition the missing word was the direct object of the verb (e.g. for *Which bucket did the lady wash the very large shirt in early this morning?* the probe was *The lady washed a* ___). For the other half of the critical items the missing word was the noun from the adjunct phrase (e.g. for *Which baby did the boy drop the very small toy on just after lunch?* the probe was *The boy dropped the toy on the* ___). For the filler items a variety of nouns and adjectives were probed. This task requires both comprehension and memory of the original sentences, but since a failure to

respond correctly can only be taken as evidence of a failure of memory, not comprehension, it will be referred to as a memory task. There were 12 practice items of similar composition to the experimental items.

Procedure

The sentences were presented using the same word-by-word presentation method as in Experiment 1. Memory probes were presented as whole sentences. Participants read two sentences followed by the memory probe for the first sentence, followed by the memory probe for the second sentence. Sentences and memory probes were presented in pairs because the task became too easy if each sentence was followed by its memory probe, but too difficult if the interval between sentence and probe was too great. The idea was to create a situation in which the participants felt that they had a realistic chance of being able to complete the memory probes if they read the sentences carefully and attentively. Responses to the memory probes were given orally and only scored as correct if the exact word from the original sentence was used.

Results

The performance on the memory task for the sentences in the Plausible- and Implausible-at-V conditions, and for comparison the fillers, is shown in Table 1. The scores were submitted to by-participant and by-item ANOVAs. In the by-participant ANOVA participant group (native or non-native) was a between subjects factor and condition (Plausible-at-V, Implausible-at-V, Filler) was a within-subjects factor. In the by-item analysis condition was a between-items factor, and participant group was a within-items factor. There was a significant main effect of participant group, $F1(1,66) = 17.30$, $p < 0.001$, $F2(1,59) = 61.29$, $P < 0.001$, scores for the natives being higher than for the non-natives. There was also a significant main effect of condition in the by-participants analysis, $F(2,65) = 5.41$, $p < 0.01$, but not in the by-items analysis, $F2 < 1.0$. The lack of significance in the by-items analysis is not surprising given the variety of question types, with varying difficulty, that were used within each condition (although note that the same questions were used in the Plausible- and Implausible-at-V conditions). The main effect of condition appeared to be mainly due to the reduced level of performance in the Plausible-at-V condition. Further tests showed that the Plausible-at-V condition was worse than the Implausible-at-V condition, $F1(1,66) = 10.56$, $p < 0.01$; $F2(1,30) = 1.23$, whereas the Implausible-at-V condition was not significantly different from the fillers, $F1(1,66) = 1.83$; $F2(1,44) = 1.12$. In none of the analyses were there any interactions involving participant group (all $Fs < 1.0$). The important point is that relative to performance on the fillers, the natives and non-natives performed similarly on the critical items. This suggests that the non-natives experienced no particular difficulty with the structure used in the critical items.

Table 1. *Performance on the memory task in Experiment 2 (percentage correct and standard deviation in parentheses).*

Condition	Non-native speakers	Native speakers
Plausible-at-V (n = 8)	60.6 (21.7)	76.8 (19.7)
Implausible-at-V (n = 8)	69.3 (19.3)	83.2 (13.5)
Fillers (n = 30)	66.4 (17.2)	81.6 (13.8)

Turning to the analysis of reading times, outlying reading times were identified and treated in the same way as in Experiment 1. The reading profiles for the Plausible- and Implausible-at-V conditions are shown in Figure 3. ANOVAs were performed on the data from the critical det–int–adj region by participants and items. There was a main effect of participant group, $F1(1,64) = 31.78$, $p < 0.001$; $F2(1,14) = 122.98$, $p < 0.001$, reading times being faster for the natives than the non-natives. There was no main effect of plausibility, F1 and F2 both < 1.0, and no interaction between participant group and plausibility, F1 and F2 both < 1.0.

Despite the lack of any plausibility effects in the overall analysis, there was some evidence in the native data for a divergence between the conditions at and following the post-verbal determiner. ANOVAs on the native data alone showed that the plausibility effect in the critical det–int–adj region was significant by subjects, $F1(1,33) = 5.74$, $p < 0.05$, but not by items, $F2(1,14) = 1.63$. Closer inspection of the data revealed that plausibility effects were localised in different places for different participants, the net result being a smearing and dilution of the effect over the post-verbal region. It also became apparent that participants who performed relatively well on the memory task (for critical and filler items combined) tended to show more immediate plausibility effects.[4]

In order to investigate the relationship between plausibility effects and memory task performance in the native group, the native participants were split into two almost equal groups according to their memory scores. The High Memory group scored 40/46 or better (mean = 41.8, n = 17) and the Low Memory group scored less than 40/46 (mean = 33.0, n = 18). The reading profiles for these two subgroups are shown in Figure 4.

[4] A similar pattern was found if only memory performance on the critical items was considered, but it was assumed that a more accurate assessment of subjects' general memory performance was provided by their overall score.

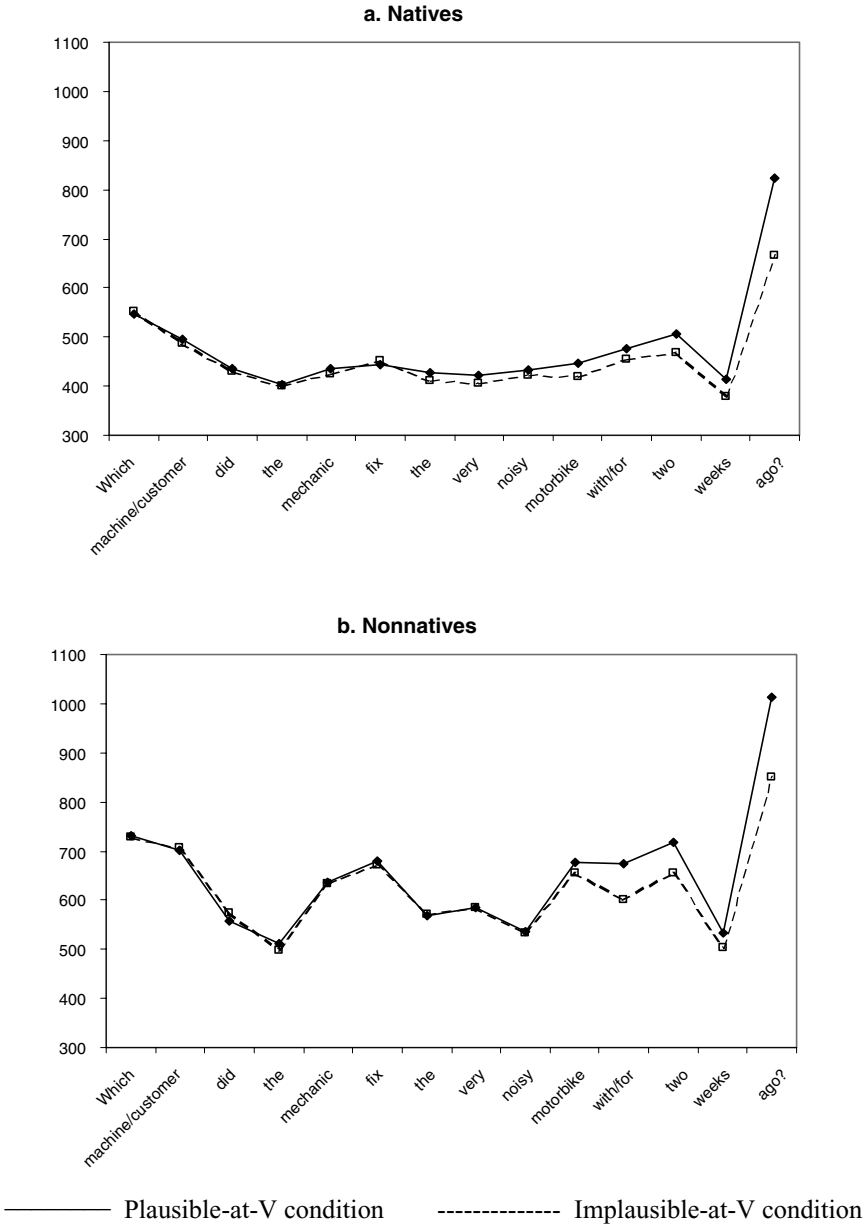

Figure 3. Experiment 2. Mean reading times (in ms).

The data from the critical det–int–adj region were submitted to an ANOVA in which Memory Group and Presentation List were between-subjects factors and Plausibility and Position were within-subjects factors. The interaction between Plausibility and Memory Group was significant by subjects, $F1(1,31) = 8.39$, $p < 0.01$, and approached significance by items, $F2(1,14) = 3.80$, $p = 0.07$.[5] As can be seen from Figure 4, the Low Memory group showed no effect of Plausibility in the det–int–adj region, whereas for the High Memory group reading time was slower in the Plausible-at-V than the Implausible-at-V condition, especially at the determiner and intensifier. An analysis of just the High Memory group data showed the effect of Plausibility to be significant in the critical det–int–adj region by subjects, $F1(1,15) = 10.99$, $p < 0.01$, and approached significance by items, $F(1,14) = 3.48$, $p = 0.08$.[6] It is also apparent from Figure 4 that the High Memory group showed a larger plausibility effect

[5] The interaction between Plausibility and Memory Group was significant by items if just the immediate post-verbal region of det–int was considered, $F2(1,14) = 21.40$, $p < 0.001$.

[6] Over the det–int region it is in fact highly significant by items, $F2(1,14) = 15.56$, $p < 0.001$.

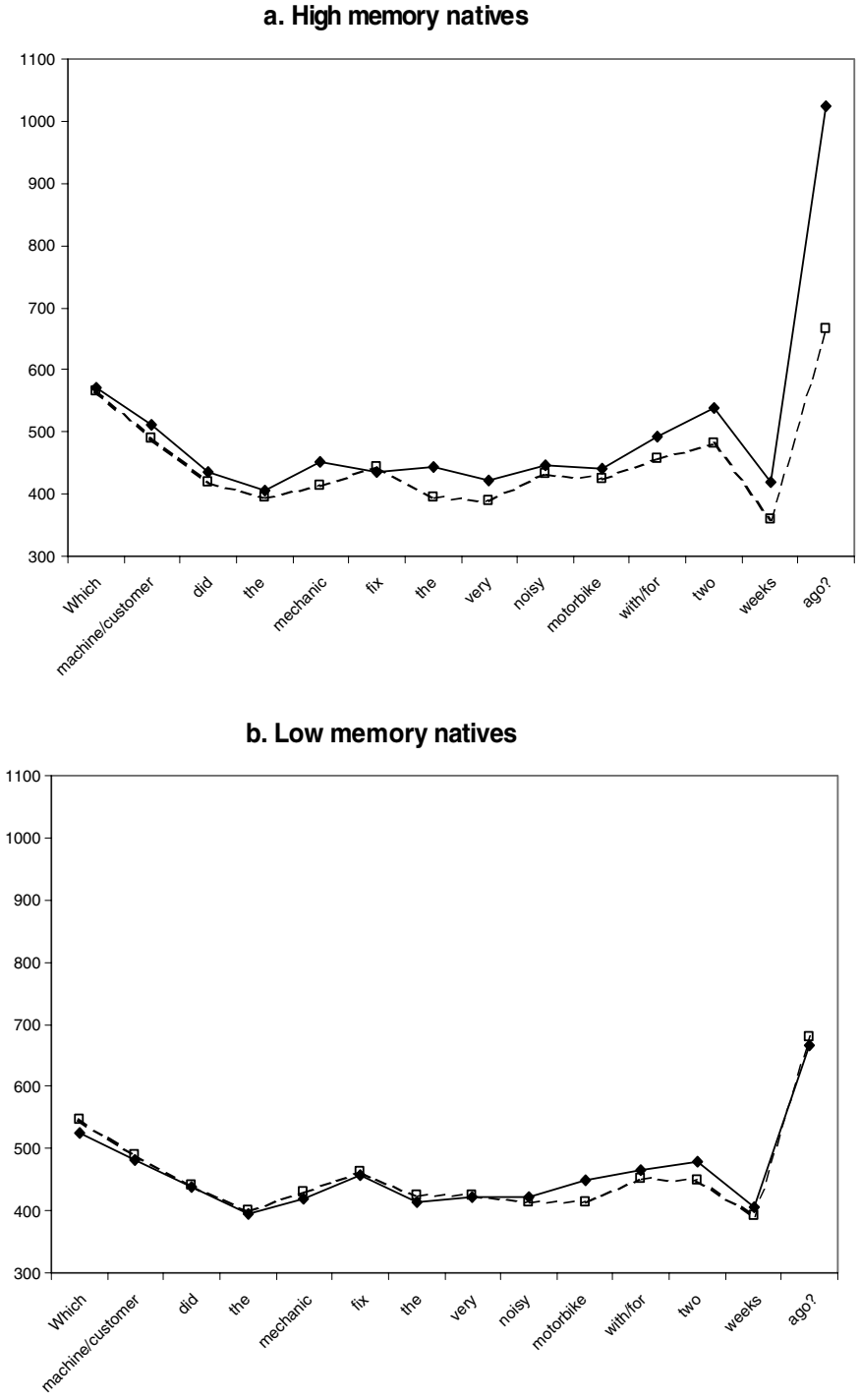

Figure 4. Experiment 2. Reading times for high versus low memory natives.

over the latter part of the sentence. ANOVAs were therefore performed on the last three word positions (e.g. *late last night*). The interaction between Memory Group and Plausibility was significant, $F1(1,31) = 4.69$, $p < 0.05$; $F2(1,14) = 8.94$, $p < 0.01$. In fact the only point at which the Low Memory group showed any sign of a plausibility effect was at the post-verbal noun. An ANOVA on the data from the Low Memory group showed that the effect of plausibility at this position was significant, $F1(1,16) = 5.96$, $p < 0.05$, $F2(1,14) = 6.71$, $p < 0.05$.

Given the large differences between the Low and High Memory groups in terms of their reading profile it is interesting that the two groups did not differ in overall reading speed for the critical sentences (463 ms

and 482 ms for the Low and High Comprehension groups respectively, $t = 0.55$).

Turning to the non-natives, Figure 3 shows that there was no effect of plausibility in the critical det–int–adj region, F1 and F2 both < 1.0. Nor was there an effect at the post-verbal noun, F1 and F2 both < 1.0. In fact it was not until the preposition that a plausibility effect emerged, $F1(1,31) = 9.61$, $p < 0.01$; $F2(1,14) = 8.35$, $p < 0.05$. However, it is important to note that some of the sentences used different prepositions in the two conditions and that the prepositions in the Plausible-at-V condition were longer than those in the Implausible-at-V condition (3.44 and 2.94 letters respectively). It was therefore necessary to evaluate whether the difference between conditions in reading time at the preposition could be due to the difference in length of the prepositions. To do this, the general relationship between reading time and word length in the non-native group was examined by calculating the linear regression equation relating the mean reading time for each word in each sentence (as derived from the items' analysis) and word length. The correlation was $r = 0.347$, $p < 0.001$. The regression equation was then used to calculate the predicted reading time for the preposition in each item. ANOVAs were then conducted on the differences between predicted and observed values in the subjects and items analyses. The effect of plausibility at the preposition was numerically reduced compared to the observed reading times (56 ms versus 75 ms) but was still significant, $F1(1,31) = 5.47$, $p < 0.05$; $F2(1,14) = 6.04$, $p < 0.05$.

There was also a plausibility effect over the remaining three words of the sentence, where the same words were used in both conditions. An ANOVA showed that the main effect of Plausibility was significant, $F1(1,31) = 10.42$, $p < 0.01$; $F2(1,14) = 5.11$, $p < 0.05$.

In order to investigate the relevance of memory task performance the participants were again split into two roughly equal groups according to their overall performance on the memory task. The High Memory group scored 33/46 or better (mean = 36.1, $n = 15$), and the Low Memory group scored less than 33/46 (mean = 25.4, $n = 18$). The reading profiles for these two groups are shown in Figure 5. Although the plausibility effects in the det–int–adj region appear to differ for the two groups, and pattern in a similar way to the natives, the interaction between Plausibility and Memory Group was far from significant, F1 and F2 both < 1.0. At the preposition, only the high memory group showed an effect of plausibility when the residual reading times were analysed, $F1(1,14) = 5.43$, $p < 0.05$; $F2(1,14) = 5.43$, $p < 0.05$, but the interaction between Memory Group and Plausibility was not significant, $F1(1,29) = 2.49$; $F2(1,14) = 2.43$. However, there was an interaction between Plausibility and Memory Group over the last three words of the sentence, $F1(1,29) = 9.42$, $p < 0.01$; $F2(1,14) = 6.38$, $p < 0.05$, the plausibility effect being larger for the high memory group. This is the same pattern that was obtained for the natives, although note that different criteria were used for dividing the participants into groups.

In summary, only the high memory native group showed any evidence for greater processing difficulty in the immediate post-verbal region in the Plausible-at-V condition. The low memory natives showed some evidence of an effect of argument competition at the noun, although this did not persist until the end of the sentence. For the non-natives, only the high memory group showed any plausibility effects, but these were delayed until the preposition. This may have reflected a slightly delayed effect of argument competition.

Two further issues need to be examined in relation to these data. The first concerns the reading profile in the critical post-verbal region in the non-natives. Was there a filled gap effect that was not sensitive to plausibility, or was there simply no filled gap effect at all? If the first were true, then it could be claimed that participants were interpreting the sentences at a structural level (because they had postulated a gap after the verb) but not at the level of the situation model. If the latter were true, then it could not be claimed that the participants were even interpreting the sentences at the structural level. This issue can be addressed by comparing the non-native reading profiles over the post-verbal region in Experiments 1 and 2, since in the former there was clearly a filled-gap effect in the Plausible-at-V condition. Figure 6 shows the reading profiles for the region from the verb to the head of the post-verbal noun phrase (e.g. *fix the very noisy motorbike*). The Chinese and Romance data from Experiment 1 have been combined into one single non-native group.

The data were submitted to an ANOVA in which Experiment and Presentation List were between-subjects factors, and Plausibility and Position were within-subjects factors. An analysis by items was also performed in which Item Group was a between-items factor, and Experiment, Plausibilty and Position were within-items factors. There was no main effect of Experiment, F1 and F2 < 1.0, indicating that the overall reading times were similar. There was an interaction between Experiment and Plausibility, $F(1, 81) = 9.26$, $p < 0.01$; $F2(1,14) = 21.24$, $p < 0.001$, reflecting the fact that the plausibility effect was larger in Experiment 1 than Experiment 2. There was also a three-way interaction between Experiment, Plausibility and Position, $F1(4, 78) = 3.36$, $p < 0.05$; $F2(4, 11) = 7.77$, $p < 0.01$. As can be seen from Figure 6, in Experiment 1 the reading times in the Plausible-at-V and Implausible-at-V conditions diverge at the intensifier and noun, whereas in Experiment 2 the reading times are equivalent at each position. For present purposes, however, what is most informative is how the reading times in Experiment 2 pattern with those in Experiment 1. Over the det–int sequence they pattern with

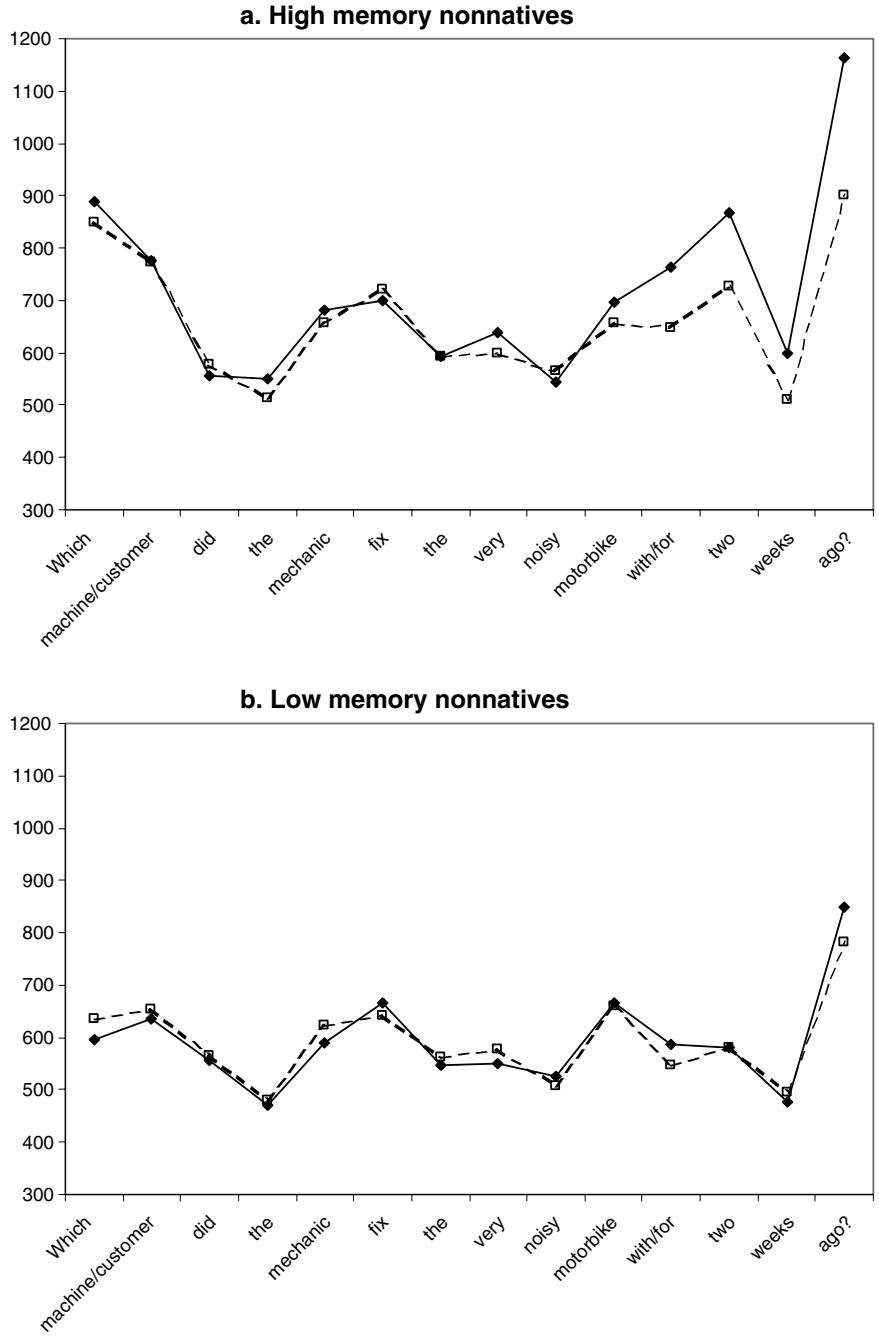

Figure 5. Experiment 2. Reading times for high versus low memory non-natives.

the Plausible-at-V condition of Experiment 1 (showing a similar rise), whilst over the adj–N sequence they pattern with the Implausible-at-V condition (all three conditions showing a lower rise than in the Plausible-at-V condition of Experiment 1). To examine the former, the data for the det–int region from the Implausible-at-V conditions were analysed. The interaction between Position and Experiment was significant, $F1(1,81) = 4.72$, $p < 0.05$; $F2(1,14) = 10.98$, $p < 0.01$. Whereas in Experiment 1 reading time decreased from the determiner to the intensifier, in Experiment 2 it increased. This pattern suggests that there was a filled-gap effect in this region in both conditions of Experiment 2 and that the participants were performing a structural analysis of the filler–gap relation. The data for the adj–N region from the Plausible-at-V conditions were also analysed. The interaction between Position and Experiment was significant, $F1(1, 81) = 4.22$, $p < 0.05$; $F2(1,14) = 24.69$, $p < 0.001$. The increase in reading time from the adjective to the noun was greater in Experiment 1 than in Experiment 2. This

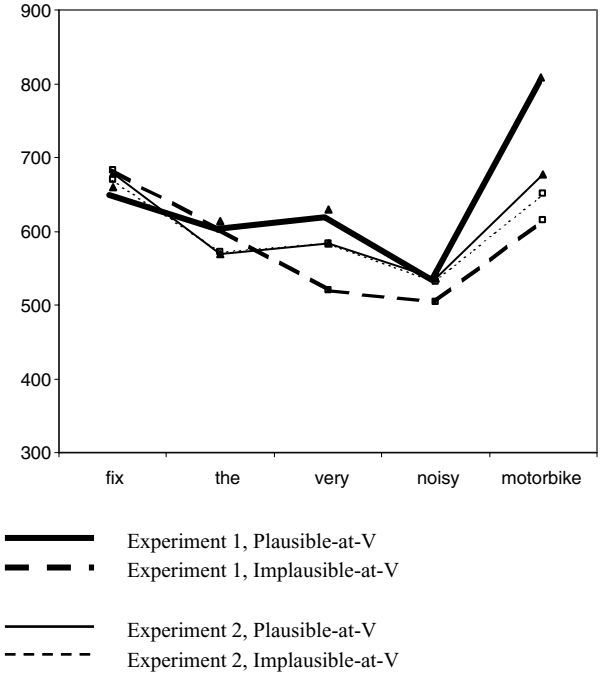

Figure 6. Non-native reading times in Experiments 1 and 2.

is consistent with the lack of a plausibility effect over the det–int–adj region. If the plausibility of the filler–gap relation were not computed, then argument competition at the post-verbal noun would be less intense than in the Plausible-at-V condition of Experiment 1, and more like that in the Implausible-at-V condition.

An equivalent analysis was performed on the native data from Experiments 1 and 2. There was a main effect of Experiment, $F1(1,48) = 9.07$, $p < 0.01$; $F2(1, 14) = 88.13$, $p < 0.001$, reading times being generally faster in Experiment 2 than Experiment 1. There was an interaction between Experiment and Plausibility on the subjects analysis, $F1(1,48) = 13.16$, $p < 0.001$, but not on the items analysis, $F2(1,14) = 2.00$. The plausibility effect was numerically greater in Experiment 1 than Experiment 2. But there was no interaction between Experiment, Plausibility, and Position on either analysis, $F1(4,45) = 1.39$; $F2(4,11) = 1.52$. The pattern of plausibility effects did not differ significantly across the two experiments.

The second issue that needs to be addressed is the source of the large plausibility effects over the latter part of the sentences, at least for the high memory natives and non-natives. These could reflect sustained argument competition as readers in the Plausible-at-V condition find it difficult to maintain a stable representation of the sentence when there are two arguments that can plausibly fulfil the same role. But these effects could also simply reflect lower levels of perceived plausibility of the Plausible-at-V sentences. For example, participants may have regarded a mechanic repairing a motorbike with a machine (Plausible-at-V condition) as less plausible than a mechanic repairing a motorbike for a customer (Implausible-at-V condition). A rating task was therefore carried out on the critical materials in which 16 advanced non-native speakers of English with a variety of L1s were asked to rate the plausibility of the events corresponding to both versions of the critical sentences. The critical sentences were transformed into event descriptions, such as *A mechanic repairing a motorbike with a machine*, and presented in random order with the constraint that the Plausible- and Implausible-at-V versions of the same item occurred in different halves of the list. A seven-point rating scale was used on which 7 signified a highly plausible event, and a 1 signified a highly implausible event. The mean rating for the events that were derived from the Plausible-at-V sentences was 5.10, and for the events derived from the Implausible-at-V sentences it was 5.14. The difference between these was not significant. Therefore it appears that the differences in reading time between these conditions over the latter part of the sentences could not have been due to the relative plausibility of the overall events described by the sentences. It seems likely, therefore, that these effects reflect sustained argument competition.

Discussion

This experiment removed the requirement to make on-line plausibility judgements. Instead participants read the sentences word by word and their memory was probed after every pair of sentences. Under these conditions the plausibility effect in the critical post-verbal region was not significant when the native and non-native groups were combined. Although, numerically, there was a small plausibility effect in the native group, this was only significant in the analysis by participants, and there was no interaction with the non-native group. However, closer inspection of the native data revealed that the plausibility effects in the det–int–adj region were significantly greater for that half of the sample with the best memory performance. In contrast, even for the half of the non-native sample with the best memory performance there was no evidence for a plausibility effect in the det–int–adj region. The variability between native participants explains why there was no interaction between participant group and plausibility in the overall analysis. It was not the case that plausibility effects were obtained for natives but not for non-natives. Rather, plausibility effects were the exception rather than the rule in this experiment, and appeared to be most likely in the natives who performed best on the memory task.

Let us first address the issue of the absence of a plausibility effect over the entire post-verbal noun phrase

in the non-native participants, regardless of memory performance. Was this because readers simply did not hypothesise a gap at the verb, and so there was no filled-gap effect, or because they did hypothesise a gap but there was no effect of plausibility on the size of the filled-gap effect? The comparison between the reading profiles of Experiments 1 and 2 suggested the latter. The reason for this is that over the critical region, the reading times in both conditions of Experiment 2 seem to pattern with the Plausible-at-V condition of Experiment 1, which we know was showing a filled-gap effect. If there were no filled-gap effect in Experiment 2, then one would expect reading times in the critical region to be relatively fast, and more like those in the Implausible-at-V condition of Experiment 1. This was clearly not the case. Readers were encountering some processing difficulty in the critical region in both conditions of Experiment 2, suggesting that a gap had been postulated at the verb and readers were encountering difficulty with the subsequent det–int–adj sequence. The lack of a plausibility effect in this region shows that although a gap was postulated, its plausibility was either (a) not evaluated, or (b) plausibility information did not immediately affect subsequent processing. Option (a) appears to apply to the low memory group of non-natives because they did not show plausibility effects anywhere in the sentences, but it cannot be true of the high memory group because they showed clear differences between the Plausible-at-V and Implausible-at-V conditions at and following the preposition. The rating task suggested that this effect was not because the sentences in the two conditions differed in overall plausibility. Instead it would appear to reflect the plausibility of the filler-as-direct-object analysis at the verb.

This brings us to the second possibility, (b), that plausibility was not immediately affecting processing in the high memory non-natives. One reason for this may be that computation of plausibility was simply delayed. The plausibility of the filler as direct object was not computed sufficiently quickly for it to influence the reader's level of commitment to one of the readings before structural cues triggered reanalysis in any case. When the post-verbal noun was encountered its plausibility was not computed sufficiently quickly for argument competition to be a problem at that point, this effect being delayed until the following word. In other words, only for the natives was the computation of plausibility time-locked to the words as they appeared on the screen, even though for all groups reading was self-paced.

Another possible explanation is that the non-native group only updated their interpretation of the filler when they encountered a relevant gap in the input. At the verb the filler would be immediately interpreted as the direct object and theme, regardless of plausibility. This analysis was only modified when the next gap was encountered, at the preposition, where it became clear that the filler performed an adjunct role. Arriving at this analysis would have been harder when the initial analysis was plausible. In this view there was no actual delay in the computation of plausibility. Rather the syntactic analysis of the input did not cause incremental changes in the situation model representation. Thus, the syntactic cues carried by the det–int–adj–N sequence signalled a post-verbal noun phrase, but this information did not influence the interpretation of the filler as the direct object until an alternative role for that filler was forced by the preposition.

Seen from the latter perspective, the results for the natives and the non-natives alike can be seen as presenting a cline of incrementality. The high memory natives show the highest degree of incrementality of interpretation, there being plausibility effects apparent from the post-verbal determiner. The low memory natives show the next level of incrementality, displaying plausibility effects at the post-verbal noun, and the high memory non-natives the next level, displaying plausibility effects from the preposition. This cline can be regarded in terms of the amount of syntactic information required to force a reanalysis of the role of the filler. Clearly this variation cuts across the nativeness distinction, there being varying degrees of incrementality even within the native group.

Of course it is not surprising that there should be a general relationship between memory task performance and plausibility effects. Participants who engaged in less elaboration of the sentence meanings, and did not construct rich situation model representations, would be expected to show poorer memory, as the classic studies on depth of processing demonstrated (Hyde and Jenkins, 1969; Bower and Winzenz, 1970). This might explain why the low memory natives and non-natives showed the least evidence of plausibility effects, especially towards the end of the sentence. But it does not tell us why there was variation in incrementality amongst participants who showed at least some evidence of plausibility effects.

One possibility is that reading speed determined the immediacy with which semantic information was computed. Slow careful reading would be associated with more incremental updating of the situation model or thematic representation. This can be immediately ruled out because the non-native groups in Experiment 1 were reading at almost exactly the same speed as the non-natives in Experiment 2, especially in the region between the verb and the preposition, where reanalysis is assumed to take place. In addition, the high and low memory sub-groups of natives in Experiment 2 were reading at the same speed. In all of these cases, differential sensitivity to plausibility was apparent in groups that had the same reading speed.

In some cases the cause may have been merely motivational. Some participants simply may not have

allocated sufficient resources to the task to construct situation models incrementally as they read. In other cases, they may not have been able to do so because of relatively high demands of lower level decoding processes (Perfetti, 1985). The latter seems particularly likely in the case of non-native readers. For example, people tend to perform worse on reading span tasks (tasks that require simultaneous processing and storage of verbal material) in their L2 than their L1 (H. C. Walter, 2000; C. Walter, 2004). This suggests that the increased demands of L2 processing may drain "working memory capacity", as defined as a common resource pool underlying storage and processing of information (Just and Carpenter, 1992). In research on natives it has been suggested that individuals with relatively low working memory capacity show less of an effect of plausibility constraints on attachment decisions during on-line syntactic parsing (Just and Carpenter, 1992), although this result has recently been contested (Clifton et al., 2003). Other researchers have argued that low working memory individuals show a more general reduction in sensitivity to plausibility in on-line processing (Pearlmutter and MacDonald, 1995). Poor L1 comprehenders have been shown to be less able to detect anomalies (Hannon and Daneman, 2004), and people with low working memory in the L2 are less able to detect semantic anomalies when reading stories in the L2 (C. Walter, 2004). It is not clear whether such effects can be construed as due to capacity limitations as such or to less efficient access of semantic information during on-line processing (cf. Just and Carpenter, 1992; Pearlmutter and MacDonald, 1995), but the consensus appears to be that at lower levels of comprehension skill, there is a reduction in sensitivity to plausibility. To the extent that L2 readers' working memory capacity is drained by increased processing demands, they might show similar effects when reading in the L2.

However, none of the above approaches to explaining variability in plausibility effects consider the impact of the nature of the task. In the present case, non-natives with a similar level of proficiency, who were reading the sentences at the same speed, showed plausibility effects on reanalysis in Experiment 1, but not in Experiment 2. Even if the results of Experiment 2 are explained in terms of limited working memory or semantic access efficiency, it has to be noted that similar individuals did show plausibility effects on reanalysis when the task demanded incremental semantic processing. Therefore, it appears to be possible to overcome processing limitations under the appropriate task conditions.

A final issue that needs to be considered is the continued lack of an effect of plausibility at the verb. Experiment 2 rules out a task-based explanation. A clue to an alternative account comes from a study by Pickering and Traxler (2001). They compared sentences like the following:

(10) That's the event that the coach persuaded the pupils to go and watch.

(11) That's the diver that the coach persuaded the pupils to go and watch.

There was no difference between the conditions at *persuaded*, despite the fact that it is not plausible to persuade an event. But there was greater processing difficulty in (11) at *the pupils* indicating that the plausibility of the initial filler–gap assignment was having an effect on reanalysis difficulty. Pickering and Traxler (2001) argue that the reason for the lack of an effect at the verb is that all of their verbs could potentially take a sentential complement. On reading *persuade* the participants were able to weigh up the probability of the filler being the direct object or of it occupying a role in the complement clause. The reader's syntactic expectations were modified accordingly, leading to no processing difficulty over the post-verbal noun phrase *the pupils* in (10), but difficulty in (11). Similarly, Boland et al. (1995) found low rates of stop making sense decisions at verbs in sentences equivalent to (10).

The present results conform to the pattern found by Pickering and Traxler (2001), the difference being, of course, that the verbs did not take sentential complements. Some of them were potentially ditransitive (these being *buy*, *push*, *kill*, *build* and *cook*), but in only two cases could the filler have plausibly been interpreted as an indirect object (see footnote 1). It would have to be assumed, therefore, that readers were able to evaluate the potential for the filler to perform an adjunct role on the basis of their knowledge of the world, rather than their knowledge of the subcategorisation frames of verbs. For example, in Experiment 1, on reading the fragment *Which customer did the mechanic repair . . .*, readers either considered only the possibility that the mechanic repaired a customer, in which case they made a stop making sense decision, or else they considered the possibility that the customer performed an adjunct role, e.g. the mechanic repaired something for/in front of/with a customer. In this case they would not have slowed down at the verb, and would have been anticipating a post-verbal noun phrase. In Experiment 2, given that there were no implausible sentences, the possibility that the mechanic repaired the customer would have been immediately rejected, leaving only the possibility of an adjunct role. What is perhaps surprising in this view is that, at least when demanded by the task, non-native readers were as able to carry out these computations as effectively as native readers.[7]

[7] It is notable that in the studies by Pickering and colleagues (Traxler & Pickering, 1996; Pickering & Traxler, 1998; Pickering & Traxler, 2003) there is actually no direct evidence for slower reading at the verb as such. This is because reading time is reported over the verb and the following word, e.g. *wrote unceasingly* in (1) and (2). It is

Conclusion

The present study explored plausibility effects in native and non-native sentence processing. One issue was whether such effects are delayed in non-native readers, or whether there is some fundamental difference in the way in which plausibility impacts upon performance, say, via argument competition rather than via a direct interaction with syntactic processing. Experiment 1 provided strong evidence against the notion of argument competition because for the non-native groups plausibility effects were evident in the disambiguating region even prior to the post-verbal noun. But the non-natives provided no evidence for delayed computation of plausibility either. Both natives and non-natives, regardless of the nature of the L1, appeared to immediately compute the plausibility of the potential filler–gap relationship at the first verb, with consequences for processing of the subsequent noun phrase. The second issue was that of task-dependence. Indeed, very different results were obtained in Experiment 2 where the task was to read the sentences in order to answer comprehension/memory questions. There was variability in the size and location of plausibility effects amongst both native and non-native groups suggesting that incrementality of interpretation depends upon general cognitive factors, such as working memory or motivation, rather than whether the person is reading in their L1 or L2. When the task demanded clear attention to meaning, non-natives produced native-like plausibility effects, but without this compulsion, the effects were delayed, although not eliminated.

The idea that retrieval of semantic information is less efficient from L2 words is of course familiar from the literature on L2 lexical processing. The Revised Hierarchical Model of bilingual memory (Kroll and Stewart, 1994) explains asymmetries in translation performance, as well as other phenomena such as reduced semantic blocking effects in L2 in terms of the reduced "strength" of links between L2 words and conceptual representations. In this view, bilinguals rely on direct lexical, translational links from L2 to L1 words when performing translation tasks, especially at lower levels of proficiency (Chen and Leung, 1989). Others argue that L2 to L1 translation is not immune from conceptual influences (La Heij, Hooglander, Kerling and van der Veleden, 1996), whilst at the same time accepting that semantic processing of L2 words is less efficient than for L1 words (La Heij et al., 1996; Talamas, Kroll and Dufour, 1999). It has also been suggested that semantic access from L2 words is particularly task-dependent. Masked translation priming effects can be obtained from L1 to L2 but not from L2 to L1. This cannot be because L2 words are simply not recognised under masking conditions because L2-to-L1 masked translation priming can be obtained when the target task is changed to semantic categorisation (Grainger and Frenck-Mestre, 1998; Finkbeiner, Forster, Nicol, and Nakamura, 2004). Assuming that even priming between translation equivalents is at least partially mediated by semantic representations (Williams, 1994), these studies suggest that semantic processing of L2 words is task-dependent. This may be because of reduced fluency of access, or, as argued by Finkbeiner et al. (2004) it may be because of differences in the underlying representation of the meanings of L2 words. Given these kinds of findings it should not be surprising that the computation of plausibility during sentence processing should be less efficient and more task-dependent in L1 than L2, although clearly there is much more that needs to be done in working out exactly how it is that these inefficiencies produce variations in the incrementality of interpretation.

There is a strand of research from both ERP (event-related potential) and behavioural studies that suggests that native and non-native sentence processing is similar along the semantic dimension, but dissimilar along the syntactic dimension (Hahne and Friederici, 2001; Felser, Roberts, Gross and Marinis, 2003; Sanders and Neville, 2003; Clahsen and Felser, in press). With regard to the latter, the present results suggest that, even if the underlying processing of syntactic information differs in native and non-natives readers, the behavioural consequences, as measured by garden-path and reanalysis effects on reading times, are essentially similar. Of course there may be areas of syntax where behavioural differences could be exposed (Marinis, Roberts, Felser, and Clahsen, 2005), but it is striking how, for the majority of structures so far examined, non-native processing is remarkably native-like. This is the case even in areas where the L1 and L2 are very different with regard to the target structure (as for the Chinese participants in Experiment 1). With regard to semantic processing, the present results point to similarities between natives and non-natives in that both groups showed variations in incrementality depending on task demands and individual variation in memory performance. To the extent that natives and non-natives are seen as falling on the same cline of incrementality, then their underlying semantic processing can be regarded as essentially similar, and subject to the same sources of variability as natives. Having said this, even if the underlying processing of semantic information is similar in native and non-native

not clear therefore whether there was an immediate effect of filler plausibility at the verb. In contrast, in the present experiments, and indeed in Pickering & Traxler (2001), disambiguating information, in the form of the definite article, occurred immediately after the verb and only the verb was regarded as the critical segment for analysis. Of course, in these experiments the following word provided immediate disambiguating information, ruling out the direct object analysis. In the Pickering & Traxler (1996, 2003) experiments the post-verbal information was not inconsistent with the direct object analysis, and so slower reading time in the Implausible-at-V condition would be expected. Therefore, it is still unclear whether implausible filler–gap assignments cause a slow-down in processing at the verb.

populations, as revealed by ERP responses to semantic anomalies for example, non-native semantic processing appears to be particularly task- and person-dependent, potentially subject to capacity constraints in the individual, or at least to strategic allocation of resources within a task. Thus, it is likely that native and non-native groups will differ in terms of their ability to perform semantic processing, even if it is carried out in essentially the same way in all populations.

From a theoretical perspective it is clearly important and interesting to know whether underlying syntactic processes are native-like. But at the same time it is important to emphasise that, despite any such underlying differences, non-natives can behave in a remarkably native-like way, especially when the relevant kind of processing is encouraged by the task.

Appendix

Plausible-at-V version
Implausible-at-V version

Which car did the man buy the really expensive radio for two months ago?
Which friend did the man buy the really expensive radio for two months ago?

Which girl did the man push the very nice bike into late last night?
Which river did the man push the very nice bike into late last night?

Which machine did the mechanic fix the really noisy motorbike with two weeks ago?
Which customer did the mechanic fix the really noisy motorbike for two weeks ago?

Which parcel did the secretary find the very large bomb in early this morning?
Which floor did the secretary find the very large bomb on early this morning?

Which relative did the farmer kill the very old chicken for two weeks ago?
Which stick did the farmer kill the very old chicken with two weeks ago?

Which ladder did the man repair the very long roof with during the holidays?
Which friend did the man repair the very long roof for during the holidays?

Which businessman did the gangster hide the very rich woman for late last night?
Which cave did the gangster hide the very rich woman in late last night?
Which dog did the farmer chase the very lively sheep with early this morning?
Which hill did the farmer chase the very lively sheep up early this morning?
Which station did the architect build the really large hotel beside during the summer?
Which mountain did the architect build the really large hotel on during the summer?

Which meal did the chef cook the really tasty meat for during the afternoon?
Which pot did the chef cook the really tasty meat in during the afternoon?

Which bucket did the lady wash the very large shirt in early this morning?
Which soap did the lady wash the very large shirt with early this morning?

Which woman did the doctor examine the very sick child for early this morning?
Which lab did the doctor examine the very sick child in early this morning?

Which baby did the boy drop the very small toy on just after lunch?
Which hole did the boy drop the very small toy in just after lunch?

Which lorry did the thief crash the very heavy car into late last night?
Which wall did the thief crash the very heavy car into late last night?

Which toy did the boy break the brand new window with in the afternoon?
Which stone did the boy break the brand new window with in the afternoon?

Which machine did the woman clean the recently repaired floors with last night?
Which detergent did the woman clean the recently repaired floors with last night?

References

Bachman, L. & Palmer, A. (1989). The construct validation of self-ratings of communicative language ability. *Language Testing*, 6, 14–29.

Boland, J. E., Tanenhaus, M. K., Garnsey, S. M. & Carlson, G. N. (1995). Verb argument structure in parsing and interpretation: Evidence from *wh*-questions. *Journal of Memory and Language*, 34, 774–806.

Bower, G. H. & Winzenz, D. (1970). Comparison of associative learning strategies. *Psychonomic Science*, 20, 119–120.

Chen, H. & Leung, Y. (1989). Patterns of lexical processing in a nonnative language. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15, 316–325.

Clahsen, H. & Felser, C. (in press). Grammatical processing in language learners. *Applied Psycholinguistics*.

Clifton, C., Traxler, M. J., Mohamed, M. T., Williams, R. S., Morris, R. K. & Rayner, K. (2003). The use of thematic role information in parsing: Syntactic processing autonomy revisited. *Journal of Memory and Language*, 49, 317–334.

Felser, C., Roberts, L., Gross, R. & Marinis, T. (2003). The processing of ambiguous sentences by first and second language learners of English. *Applied Psycholinguistics*, 24, 453–489.

Fender, M. (2001). A review of L1 and L2/ESL word integration skills and the nature of L2/ESL word integration development involved in lower-level text processing. *Language Learning*, 51, 319–396.

Finkbeiner, M., Forster, K., Nicol, J. & Nakamura, K. (2004). The role of polysemy in masked semantic and translation priming. *Journal of Memory and Language*, 51 (1), 1–22.

Frazier, L. (1987). Sentence processing: A tutorial review. In M. Coltheart (ed.), *Attention and performance XII: The psychology of reading*, pp. 559–586. Hove: Lawrence Erlbaum Associates.

Frazier, L. & Clifton, C. (1989). Identifying gaps in English sentences. *Language and Cognitive Processes*, 4, 93–126.

Grainger, J. & Frenck-Mestre, C. (1998). Masked priming by translation equivalents in bilinguals. *Language and Cognitive Processes*, 13, 601–623.

Hahne, A. & Friederici, A. D. (2001). Processing a second language: Late learners' comprehension mechanisms as revealed by event-related brain potentials. *Bilingualism: Language and Cognition*, 4, 123–141.

Hannon, B. & Daneman, M. (2004). Shallow semantic processing of text: An individual differences account. *Discourse Processes*, 37, 187–204.

Hoover, M. L. & Dwivedi, V. D. (1998). Syntactic processing in skilled bilinguals. *Language Learning*, 48 (1), 1–29.

Hyde, T. S. & Jenkins, J. J. (1969). Differential effects of incidental tasks on the organization of recall of a list of highly associated words. *Journal of Experimental Psychology*, 83, 472–481.

Juffs, A. (2004). Representation, processing and working memory in a second language. *Transactions of the Philological Society*, 102, 199–225.

Juffs, A. & Harrington, M. (1996). Garden path sentences and error data in second language sentence processing. *Language Learning*, 46 (2), 283–326.

Just, M. A. & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Boston, MA: Allyn & Bacon, Inc.

Just, M. A. & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review*, 99, 122–149.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95, 163–182.

Kroll, J. F. & Stewart, E. (1994). Category Interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, 33 (2), 149–174.

La Heij, W., Hooglander, A., Kerling, R. & van der Veleden, E. (1996). Nonverbal context effects in forward and backward word translation: Evidence for concept mediation. *Journal of Memory and Language*, 35, 648–665.

Marinis, T., Roberts, L., Felser, C. & Clahsen, H. (2005). Gaps in second language sentence processing. *Studies in Second Language Acquisition*, 27, 53–78.

Pearlmutter, N. J. & MacDonald, M. C. (1995). Individual differences and probabilistic constraints in syntactic ambiguity resolution. *Journal of Memory and Language*, 34, 521–542.

Perfetti, C. A. (1985). *Reading ability*. Oxford: Oxford University Press.

Pickering, M. J. & Barry, G. (1991). Sentence processing without empty categories. *Language and Cognitive Processes*, 6, 229–259.

Pickering, M. J. & Traxler, M. J. (1998). Plausibility and recovery from garden paths: An eye-tracking study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 24 (4), 940–961.

Pickering, M. J. & Traxler, M. J. (2000). Parsing and incremental understanding during reading. In M. W. Crocker & M. Pickering (eds.), *Architectures and mechanisms for language processing*, pp. 238–258. New York: Cambridge University Press.

Pickering, M. J. & Traxler, M. J. (2001). Strategies for processing unbounded dependencies: Lexical information and verb-argument assignment. *Journal of Experimental Psychology: Learning Memory and Cognition*, 27, 1401–1410.

Pickering, M. J. & Traxler, M. J. (2003). Evidence against the use of subcategorisation frequency in the processing of unbounded dependencies. *Language and Cognitive Processes*, 18 (4), 469–503.

Sanders, L. D. & Neville, H. J. (2003). An ERP study of continuous speech processing II: Segmentation, semantics, and syntax in non-native speakers. *Cognitive Brain Research*, 15, 214–227.

Stowe, L. (1986). Parsing *wh*-constructions: Evidence for on-line gap location. *Language and Cognitive Processes*, 1, 227–245.

Talamas, A., Kroll, J. F. & Dufour, R. (1999). From form to meaning: Stages in the acquisition of second-language vocabulary. *Bilingualism: Language and Cognition*, 2 (1), 45–58.

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

Walter, C. (2004). Transfer of reading comprehension skills to L2 is linked to mental representations of text and to L2 working memory. *Applied Linguistics*, 25, 315–339.

Walter, H. C. (2000). *The involvement of working memory in reading in a foreign language*. Ph.D. dissertation, University of Cambridge.

Williams, J. N. (1994). The relationship between word meanings in the first and second language: Evidence for a common, but restricted, semantic code. *European Journal of Cognitive Psychology*, 6, 195–220.

Williams, J. N., Möbius, P. & Kim, C. (2001). Native and non-native processing of English *wh*-questions: Parsing strategies and plausibility constraints. *Applied Psycholinguistics*, 22, 509–540.

Received May 19, 2003
Revision received May 1, 2005 and July 26, 2005
Accepted August 24, 2005