Do self-paced reading studies provide evidence for rapid syntactic adaptation?

Grusha Prasad and Tal Linzen
Johns Hopkins University

Syntactically ambiguous sentences that are disambiguated in favor of a less preferred parse are read more slowly than their unambiguous counterparts. This is called a *garden path effect*. Recent studies have found that this garden path effect decreased as participants are exposed to many such syntactically ambiguous sentences over the course of an experiment. This decrease has been interpreted as evidence for rapid syntactic adaptation — i.e. evidence that readers rapidly calibrate their expectations to a new environment in order to minimize how surprised they are when they encounter these unexpected syntactic structures (Fine, Jaeger, Farmer, & Qian, 2013). Syntactic adaptation is only one possible explanation for the observed decrease in garden-path effect, however: this decrease could also be driven by increased familiarity with the experimental paradigm (*task adaptation*), which impacts difficult sentences more than it does easy ones. The goal of this paper is to tease apart these two explanations. Using a between-group design, we demonstrate that the decrease in garden path effect is not dependent on readers' experience during the experiment. This suggests that it is unlikely to be driven primarily by syntactic adaptation. We also provide preliminary evidence that the decrease in garden path effect is driven by asymmetric effects of task adaptation. We conclude that self-paced reading studies cannot provide unambiguous evidence for rapid syntactic adaptation.

Introduction

Humans' ability to extract statistical regularities from their environment plays an important role in language acquisition and processing (Mitchell, Cuetos, Corley, & Brysbaert, 1995; Romberg & Saffran, 2010). In sentence comprehension, in particular, predictable syntactic structures are easier to process than unpredictable ones (MacDonald, Pearlmutter, & Seidenberg, 1994; Trueswell, 1996). Readers' estimates of the frequency of a syntactic structure have recently been argued to be highly reflective of recent experience. Such malleability would be expected under a rational analysis of sentence comprehension (Anderson, 1990): since the distribution of syntactic structures can vary widely across environments and contexts, readers' expectations will only be an accurate reflection the statistics of the current environment if readers can rapidly calibrate their expectations to match those statistics (Fine et al., 2013).

In line with this framework, Wells, Christiansen, Race, Acheson, and MacDonald (2009) showed that participants exposed to sentences with relative clauses over several experimental sessions found it easier to process such sentences than participants exposed to sentences with other syntactic structures. Fine et al. (2013) argued that syntactic adaptation can occur within a single experimental session. They tested this prediction by measuring the *garden-path effect* for temporarily ambiguous sentences such as (1a). The word *conducted*, which disambiguates the sentence in favor of the originally dispreferred reduced relative clause (RRC) parse, is read more slowly in (1a) than in the control sentence (1b) (MacDonald, Just, & Carpenter, 1992):

(1) a. The experienced soldiers warned about the dangers conducted the midnight raid. (RRC-disambiguated)
b. The experienced soldiers who were told about the dangers conducted the midnight raid.

Fine et al. (2013) showed that the garden path effect decreased over time, suggesting that sentences like (1a) were becoming increasingly easy to process. They interpreted this as evidence for *syntactic adaptation* — i.e. evidence that participants updated their expecations to match the statistics of the environment. This decrease in garden path effect over time has since been observed in other studies (Fine & Jaeger, 2016; Harrington Stack, James, & Watson, 2018).

While the decrease in garden path effect is consistent with syntactic adaptation, another explanation is possible: it could also be driven by increased familiarity with the experimental paradigm. In particular, previous studies that found the decrease in garden path effect over time (Fine & Jaeger, 2016; Fine et al., 2013; Harrington Stack et al., 2018) also found that as the experiment progressed, there was an overall decrease in reading times across all conditions. We will refer to this overall decrease in reading time as *task adaptation*. There is a limit to how much task adaptation we can expect: self-paced reading times reflect key presses, and those cannot fall below a certain threshold. Given such floor effects,

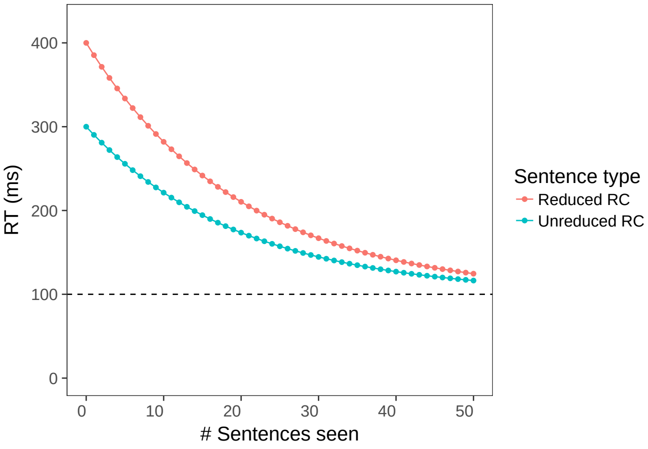


Figure 1. Illustrating the asymmetric decrease in RTs over time on simulated data. The following exponential function was used: $RT = m.e^{\frac{T}{20}} + 100$ where T is the trial number and m+100 is the RT for the sentence at $T = 0$. The dashed line at 100 ms represents the floor (i.e. the minimum possible RT for any word)

conditions that are read more slowly at the beginning of the experiment (RRC-disambiguated, (1a)) may exhibit a larger task-driven decrease in reading time than those that are read more quickly (1b). Since the garden path effect is defined as the difference between these types of sentences, floor effects in task adaptation would manifest as a decrease in the garden path effect over time. Figure 1 illustrates a hypothetical change in RTs for two conditions over time using an exponential function to approximate task adaptation; this simple simulation demonstrates that we may be able to model differences between conditions over time without appealing to syntactic adaptation.

The goal of this paper is to tease apart these two explanations using a control group. If the decrease in the garden-path effect were driven by syntactic adaptation, then we would expect a group exposed to a particular number of RRC-disambiguated sentences to show a smaller garden-path effect than a group exposed to the same number of filler sentences with other structures. By contrast, if the decrease in the garden-path effect were driven by task adaptation, then reading time will be affected only by the number of sentences to which a participant has been exposed rather than the structure of those sentences, and we would not expect to find a difference in garden-path effect between groups.

In fact, this logic was already implemented by Fine et al. (2013). In their study, the garden-path effect for a group exposed to 16 sentences with RCs (8 reduced, 8 unreduced) was moderately smaller than that for a control group exposed to 16 fillers, in line with the predictions of the syntactic adaptation account; however, this difference did not reach significance. In a highly powered follow-up study using the same design with more items and participants, Harrington Stack et al. (2018) did not find evidence for a difference in garden-path effects between the groups; they concluded that their findings do not provide evidence for syntactic adaptation. One potential criticism of the Harrington Stack et al. (2018) study is that in order to increase the number of items they repeated verbs such that the same participant saw the verb in both reduced and unreduced relative clauses. As Harrington Stack et al. acknowledge, this aspect of their design may have decreased their ability to detect adaptation effects. While existing evidence is consistent with the hypothesis that the decrease in garden-path effect might not be driven by syntactic adaptation, then, the support for this hypothesis is currently inconclusive.

We ran two experiments to further tease apart the two explanations. In Experiment 1, we replicated Fine and Jaeger (2016) (henceforth referred to as FJ16) in order to ensure that the decrease in the GPE over time is robust; to our knowledge, this is the first attempt to replicate FJ16 (Harrington Stack et al. 2018 attempted to replicate Fine et al. 2013). Across three experiments, FJ16 presented their participants with 20 sentences with reduced relative clauses (like (2a)) and 20 with unreduced relative clauses (like (2b))

(2) a. The evil genie served the golden figs went into a trance.
b. The evil genie who was served the golden figs went into a trance.

They found a decrease in the garden path effect (*GPE*) over the disambiguating region (underlined) when there was verb repetition (their Experiment 1), in the absence of verb repetition (their Experiment 2) and in the absence of any lexical overlap between sentences (their Experiment 3). We decided to replicate FJ16's Experiment 2 instead of Fine et al. (2013) both because the experiment had more items and because it was a simpler design.[1] In our Experiment 2, we tease apart the two explanations by using a between group design where during the exposure phase participants were either exposed to RRC-disambiguated sentences (*RRC-exposed group*) or to fillers (*Filler-exposed group*). The stimuli for both experiments and the scripts used to create the experiments, generate the plots and run all the analyses can be found on OSF[2].

To anticipate our results, Experiment 1 replicated the results of FJ16 in both direction and magnitude. At the same time, Experiment 2 did not find a difference in the garden-path effect between the RRC-exposed group and the filler-exposed group. Finally, exploratory analyses revealed that

[1] Fine et al. (2013) had an additional condition where they compared the GPE for RRC-disambiguated sentences with sentences that were disambiguated in favour of the main verb reading, e.g., *The experienced soldiers warned about the dangers before the midnight raid.*

[2] `https://osf.io/qd8ye/`

for both the RRC-disambiguated sentences and filler items, sentences which were read most slowly at the beginning showed the highest rate of decrease in RT over the course of the experiment. These results are consistent with the hypothesis that the decrease in the GPE over time could be completely attributed to asymmetric effects of task adaptation.

Experiment 1

Method

Participants. We recruited 80 participants via Prolific Academic, a crowdsourcing platform similar to Amazon's Mechanical Turk. All participants listed English as their first language on their profiles.

Materials. We used the same 40 critical items and 80 fillers as FJ16. Each of the critical items had a reduced form, as in (2a) above (repeated here as (3a)), and an unreduced form, as in (3b):

(3) a. The evil genie served the golden figs <u>went into a</u> trance.
b. The evil genie who was served the golden figs <u>went into a</u> trance.

All filler items had main verbs that have a different form for the past participle and the past tense. In other words, they included verbs like *woke*, which can only be interpreted as a past tense verb (the past participle would be *woken*), but not verbs like *served*, which is ambiguous between the two forms and therefore gives rise to the temporarily syntactic ambiguity illustrated in (3a).

We generated four pseudorandom orders and two lists counterbalanced for sentence type (i.e. reduced vs. unreduced) for each random order. In all the pseudorandom orders, stimuli were distributed across five blocks, where each block had eight critical items (four reduced, four unreduced) and 16 fillers. Every two critical items were separated by at least one filler, and at most two critical items of the same condition were allowed to follow each other. We generated a reversed version of each of the eight lists (four pseudorandom orders × two counterbalanced lists), for a total of 16 lists. Each participant was assigned to one of these 16 lists.

Procedure. The experiment was hosted on IbexFarm using the self-paced reading mode (Drummond, von der Malsburg, Erlewine, & Vafaie, 2016). At the beginning of every trial, each word in the sentence was replaced by a dash whose length was roughly equivalent to the length of the word. When the participant pressed the space bar, the next word in the sentence replaced the dash, and the previous word in the sentence disappeared. At the end of the sentence, participants were presented with comprehension questions, and used the keys 'z' and 'm' to respond with 'yes' and 'no' respectively. The correct answers was 'yes' half of the time. Participants were asked to fill out a brief demographic survey before the experiment, and were given three practice trials before the experiment proper started.

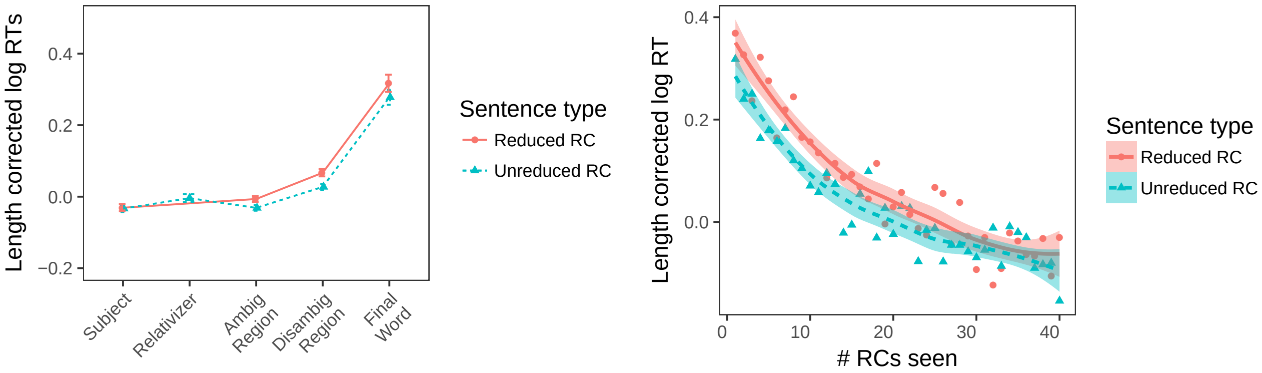
Results

Data filtering and exclusion. Six of our participants reported that English was not their first language in our demographic survey (despite the fact that we indicated on Prolific Academic that all of the workers must have English as their first language). We excluded these participants from our analyses. We then calculated the mean accuracy on fillers for all participants (excluding two fillers whose mean accuracy was lower than two standard deviations from the mean accuracy of all fillers). We excluded three participants whose accuracy was lower than 80%. Based on the data exclusion criteria in the original study, all observations with reading times lower than 100 ms or greater than 2000 ms were excluded. This led to the exclusion of 0.48% of all observations. We then excluded all the trials where participants responded incorrectly to the comprehension questions. This led to the exclusion of an additional 6.11% of our sentences (4.72% of all fillers, 10.30% of all ambiguous and 7.59% of all unambiguous)

As in the original study, in order to correct for the effects of word length on reading times, we fit a linear mixed effects model predicting log-transformed RTs from word length. Word length was centered at the mean and scaled by its standard deviation. The model also included a by-participant random intercept and slope for word length. We use the residuals of this model (henceforth referred to as *length-corrected log RTs*) in all of our analyses. We planned to exclude any participants whose mean length corrected log RTs were three standard deviations lower than the mean for all participants, but this did not apply to any of our participants. The following plots and analyses are based on the 73 remaining participants.

Statistical analysis. FJ16 divided the sentence into five regions: subject (*the experienced waitress* in (2b)), relativizer (*who was*), ambiguous region (*cooked the grilled chicken*), disambiguating region (*sent her food*) and final word (*back.*). We only analyzed the length-corrected log RTs for the disambiguating region. We fit a linear mixed effects model identical to the model in FJ16, with the following predictors:

- Sentence type: A categorical variable coded as 1 for sentences with reduced RCs and −1 for sentences with unreduced RCs. FJ16 referred to this predictor as 'ambiguity'.
- Critical item number: The number of critical items (reduced and unreduced) the participant has seen so far. FJ16 referred to this as 'item order'.
- log(Stimulus number): The natural log of the total number of sentences (critical items and fillers) the the participant has seen so far. FJ16 referred to this as



(a) (b)

Figure 2. Results of Experiment 1. (a) Length corrected log RTs for the different regions averaged over all participants and items. The error bars represent 95% CI (estimated parametrically). (b) Length corrected log RTs as a function of the number of critical items (both reduced and unreduced) averaged across all participants and items. We fit the data points with a LOESS curve.

'stimulus order'.

- Interaction of sentence type and critical item number.

Both critical item number and log stimulus order were centered around their mean. We specified the maximal random effect structure that allowed the model to converge. This included by-item and by-participant random intercepts, along with by-participant and by-item random slopes for sentence type, critical item number and the interaction between ambiguity and critical item number. There was an additional by-participant random slope for log(stimulus number). We estimated p-values for the coefficients of this model using the Satterthwaite's method in the lmerTest package in R (Kuznetsova, Brockhoff, & Christensen, 2017).

To analyze accuracy, we fit a logistic mixed effects model with fixed slopes for log stimulus number, sentence type and their interaction. The random effect structure for this model included a random intercept for participant and item, a by-item random slope for log stimulus number and by-participant random slopes for log stimulus number, sentence type and their interaction.

Results. Replicating FJ16, there was a significant GPE ($\hat{\beta} = 0.019$, $SE = 0.005$, $p \ll 0.01$) (see Figure 2a). Length corrected log RTs decreased significantly as a function of both stimulus number ($\hat{\beta} = -0.090$, $SE = 0.014$, $p \ll 0.01$) and critical item number ($\hat{\beta} = -0.003$, $SE = 0.001$, $p = 0.02$). Crucially, this decrease was more rapid for RRC-disambiguated sentences than for controls ($\hat{\beta} = -0.001$, $SE = 0.0003$, $p < 0.01$) (see Figure 2b). The coefficient of this interaction was identical to that reported by FJ16 ($\hat{\beta} = -0.001$).

Also consistent with FJ16, we found that the accuracy for sentences with reduced RCs was significantly lower than the accuracy for other sentences ($\hat{\beta} = -0.524$, $SE = 0.124$, $p \ll 0.01$). As with FJ16, we failed to find a significant decrease in accuracy over the course of the experiment ($\hat{\beta} = -0.12$, $SE = 0.086$, $p = 0.15$), suggesting that the observed effects cannot be attributed to decrease in attention over time.

Discussion

We replicated the decrease in garden-path effect over time reported by FJ16 in both direction and magnitude. This increases our confidence in the robustness of this effect. The main goal of the present paper is to test if this decrease is driven by syntactic adaptation. In our first attempt to do so, we ran a pilot study with a between-subject design, where participants were either exposed to 24 RRC-disambiguated sentences (RRC-exposed group) or 24 fillers (Filler-exposed group). The garden-path effects for both groups were compared before and after exposure. Since the RRC-exposed group was presented with more RRC-disambiguated sentences than in Experiment 1, we minimally expected to find a decrease in the garden-path effect from pre-exposure to post-exposure for this group. However, we failed to find a decrease in garden-path effect for either group.

Though the total number of RRC-disambiguated sentences was greater in our pilot study than in Experiment 1, due to limitations of the experimental design, we were able to measure the GPE over fewer items — only 16 items in the pilot study (8 reduced and 8 unreduced RCs), compared to 40 items in Experiment 1 (20 reduced and 20 unreduced RCs). During the exposure phase the RRC-exposed group was presented with only RRC-disambiguated sentences whereas the Filler-exposed group was presented with only fillers. Neither group was presented with sentences with unreduced RCs. As a consequence we could not include items in the exposure phase when measuring the GPE.

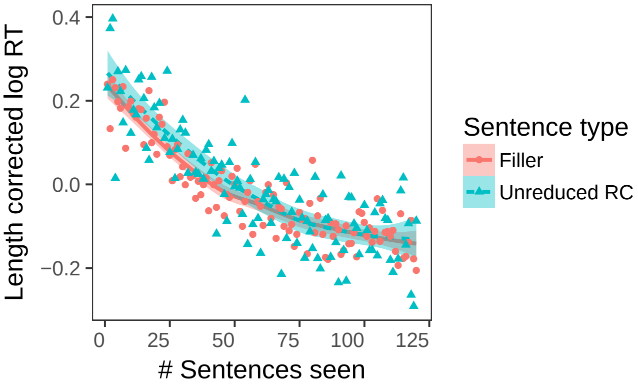


Figure 3. Length corrected log RTs over time averaged across participants and items for sentences with unreduced RCs and fillers. We fit the data points with a LOESS curve.

In Experiment 2, we modified the design of our pilot study in the following two ways, in order to be able to measure the GPE over more items. First, we did not calculate the GPE before the exposure phase. Rather, the items that were used to estimate the GPE prior to exposure were instead used to estimate the GPE post exposure. Second, we replaced all the sentences with unreduced relative clauses with RRC-disambiguated sentences. Instead of a true garden-path effect comparing reduced and unreduced sentences, we measured a "proxy-garden-path effect" by comparing the RTs for RRC-disambiguated sentences with the RTs for the fillers. Both these changes allowed us to increase the number of items we could include in our analyses by four times (compared to the pilot study).

We believe that measuring the proxy-garden path effect (*proxy-GPE*) instead of a true garden path effect is not a concern for our experiment, for two reasons. First, when we compared the length corrected log RTs for the words in the disambiguating region for sentences with unreduced RCs with words in fillers in the same sentence positions in Experiment 1, we found that they were nearly identical on average (see Figure 3). Second, and more importantly, in this experiment we are primarily interested in testing the differences between groups. Since we compare the same RRC-disambiguated sentence and filler pairs across both groups, whatever discrepancy between the true GPE and the proxy-GPE will be consistent across groups.

Experiment 2

Methods

Participants. We recruited 203 participants (three participants recruited unintentionally) via Amazon's Mechanical Turk. In order to limit the number of non-native speakers, participants were only recruited if the home address associated with their Amazon account was located in the United States.

Some of the sentences from FJ16 had verbs that have a transitivity bias—they typically occur with a noun phrase (NP) complement. Due to this bias, these sentences were essentially disambiguated before the second verb (cf. Malone & Mauner, 2018). The following sentence from F16's materials, for example, is in practice disambiguated in favour of the RRC reading at the prepositional phrase (*in the alley*), rather than the second verb (*ran*) as in some of the other items:

(4) The calico cat licked in the alley ran into the street.

To ensure that all sentences were disambiguated at the same region, we replaced 27 of FJ16's sentences with sentences constructed using optionally reflexive verbs (5a), ditransitive verbs (5b) or optionally transitive verbs without a strong transitivity bias (5c):

(5) a. The bearded man shaved two weeks ago liked his stylish new look.
b. The helpful librarian lent the frayed book took good care of it.
c. The ferocious lions attacked during the day were unable to escape the hunters.

Four of FJ16's original stimuli did not start with *the*; we added that word to these sentences for consistency. As a result, all 40 sentences had seven words before the disambiguating region (three words in the subject NP, one verb, three words in NP/PP following the verb).

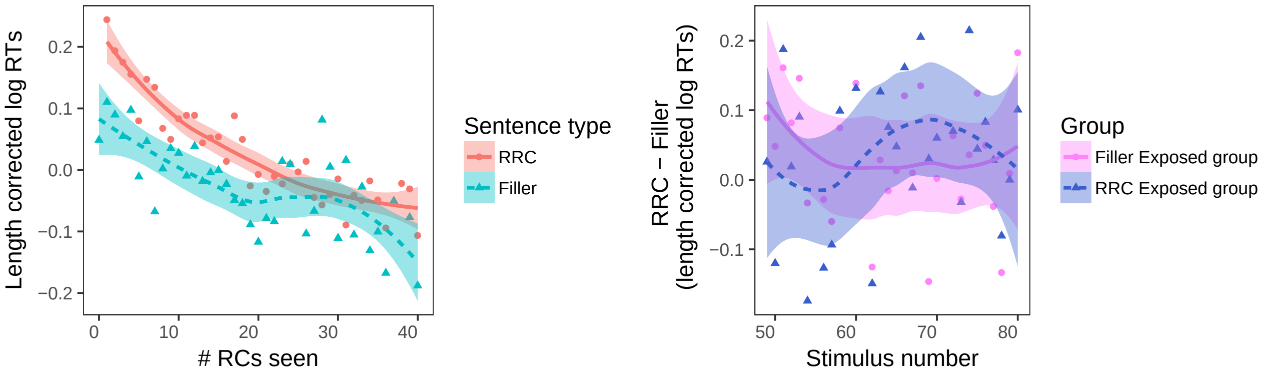
As with Experiment 1, all the stimuli were divided across five blocks, the distinction between which was not made apparent to the participants. In the exposure phase (Block 1–3), the RRC-exposed group was presented with 24 RRC-disambiguated sentences and 24 fillers, and the Filler-exposed group was presented with 48 fillers. In a subsequent test phase (Blocks 4–5), both groups were presented with 16 new RRC-disambiguated sentences and 16 new fillers.

We generated four pseudorandom orders, from which we created eight lists, one Filler-exposed list and one RRC-exposed list per random order. The pseudorandom orders were generated such that each block had eight reduced RCs, and no more than three RRCs occurred consecutively. Each participant was assigned to one of these eight lists.

Procedure. The procedure was identical to Experiment 1.

Results

Data filtering and exclusion. We used the same filtering and exclusion criteria as in Experiment 1. We excluded eight participants who were non-native speakers of English and 28 participants whose accuracy on the comprehension questions for fillers was lower than 80%. One additional par-



(a) (b)

Figure 4. Results of Experiment 2. (a) Corrected log RTs over time for the RRC-exposed group in sentence positions 8-10 averaged across participants and items as a function of the number of RRC-disambiguated sentences seen. (b) Difference between RRC and fillers in the test phase averaged across items and participants for the RRC-exposed group and Filler-exposed group. We fit the data points with a LOESS curve.

ticipant was excluded because their mean length corrected log RTs was 3 standard deviations below the mean of all participants. We excluded 0.55% of all observations because the RTs for those observations were lower than 100 ms or greater than 2000 ms. Participants responded incorrectly on 7.16% of all trials (5.82% fillers, 9.52% RRC); these trials were excluded.

Analysis 1: Change in proxy-GPE over time within the RRC-exposed group. We first analyzed the results within the RRC-exposed groups only. We fit a linear mixed effects model to the length corrected log RTs for words in sentence positions 8–10 (the sentence positions that correspond to the disambiguating region for the RRC-disambiguated sentences; *critical region*). This model had the same predictors as the model in Experiment 1: slopes for sentence type, critical item number, log stimulus number and the interaction of critical item number and sentence type. Sentence type was coded as 1 for RRC-disambiguated sentences and −1 for fillers. The critical item number and stimulus number were not only centered around their mean (like in Experiment 1) but also scaled by their standard deviation (to facilitate model convergence). This was true for all following analyses. We specified the maximal random effect structure that allowed the model to converge, which included a by-item and by-participant random intercept; by-participant slopes for sentence type, RRC number, their interaction and stimulus number; by-item slopes for RRC number and stimulus number.

The proxy-GPE was of almost the same magnitude as in Experiment 1, but was not significant ($\hat{\beta}$ = 0.018, *SE* = 0.010, p = 0.07). As in Experiment 1 and FJ16, RTs in the disambiguating region significantly decreased as a function of both stimulus number ($\hat{\beta}$ = 0.041, *SE* = 0.011, $p < 0.001$) and critical item number ($\hat{\beta}$ = −0.027, *SE* = 0.012, p = 0.03). There was also a decrease in proxy-GPE over time ($\hat{\beta}$ = −0.011, *SE* = 0.003, $p < 0.001$) (see Figure 4a).

To summarize, we observed a decrease in proxy-GPE over time for participants in the RRC-exposed group which was comparable to the change in the GPE over time in Experiment 1.

Analysis 2: Between-group differences in test phase proxy-GPE. We fit a linear mixed effects model to the length corrected log RTs in the test phase (Block 4–5) averaged across words in the critical region, for both the RRC-exposed and Filler-exposed groups. The predictors for this model included sentence type, group, log stimulus number along with the two-way and three-way interactions between these predictors. We included stimulus number as a predictor in order to capture any adaptation over the course of the test phase. As in the earlier model, sentence type was coded as 1 for RRCs and −1 for fillers. Group was coded as 1 for the RRC-exposed group and −1 for the Filler-exposed group.

Figure 4b shows the change in proxy-GPE over the course of the test phase for both groups. The patterns appear to differ across groups: the proxy-GPE for the Filler-exposed group decreased over the first few items and then stabilized, whereas the proxy-GPE for the RRC-exposed group fluctuated without a clear trend. However, this difference between the groups was not significant ($\hat{\beta}$ = 0.005, *SE* = 0.0029, p = 0.10). The only significant effect we found was an overall decrease in length corrected RTs in the critical region over the test phase ($\hat{\beta}$ = −0.016, *SE* = 0.0056, $p < 0.01$).

Participants in the Filler-exposed group also read words in the critical region in the test phase (in both fillers and RRC-disambiguated sentences) more slowly than participants in the RRC-exposed group, although this difference was not significant ($\hat{\beta}$ = −0.014, *SE* = 0.0075, p = 0.07). This pattern is consistent with previous studies (Fine et al., 2013; Harrington Stack et al., 2018). Exploratory analyses revealed

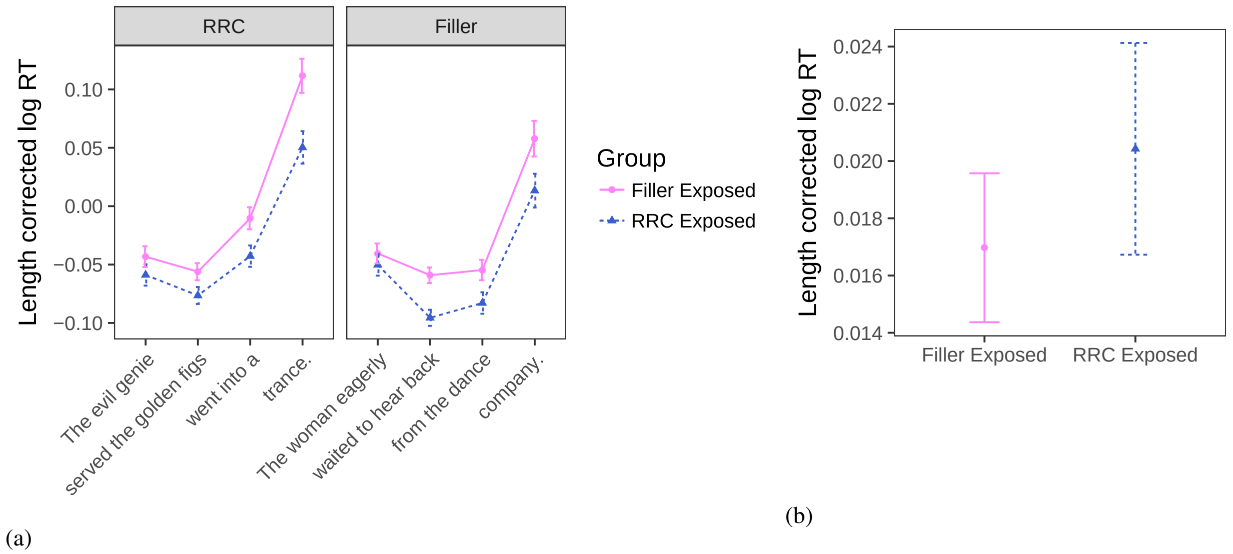


Figure 5. Results of Experiment 2. (a) Length corrected log RTs averaged across participants and items in the test phase for fillers and RRC-disambiguated sentences. (b) Length corrected log RTs averaged across the entire sentence in the exposure phase for fillers.

that this pattern was not limited to the critical region, but was consistent across the entire sentence (see Figure 5a). In the exposure phase, however, this was not true. The Filler-exposed group read sentences more rapidly on average than the RRC-exposed group (see Figure 5b). Taken together, these results suggest that participants in the Filler-exposed group started reading sentences in the test phase more slowly than participants in the RRC-exposed group. We briefly speculate about the reason for this in the General Discussion.

To summarize, we failed to find a difference in the proxy-GPE between the RRC-exposed group and the Filler-exposed group in the test phase. However, there was an overall difference in RTs between the groups — participants in the Filler-exposed group read sentences more slowly than participants in the RRC-exposed group.

Analysis 3: Change in accuracy over time. We fit a logistic mixed effects model to investigate whether there was a change in accuracy over time and if yes, whether this change differed across groups. The model had fixed slopes for sentence type, group and stimulus number along with the two-way and three-way interactions between these predictors. Sentence type and group were coded as in Analysis 2. The maximal random effects structure was specified; this included, along with random intercepts for participant and item, by-participant slopes for sentence type and stimulus number and by-item slopes for group and stimulus number.

As in Experiment 1, overall accuracy did not significantly change over time ($\hat{\beta} = 0.099$, $SE = 0.16$, $p = 0.53$). The accuracy for RRC-disambiguated sentences was significantly lower than the accuracy for fillers ($\hat{\beta} = -0.470$, $SE = 0.16$, $p < 0.01$). The accuracy for the RRC-exposed group was significantly greater than the Filler-exposed group ($\hat{\beta} = 0.317$, $SE = 0.16$, $p = 0.04$). There was also a significant three-way interaction between sentence type, group and stimulus number ($\hat{\beta} = -0.294$, $SE = 0.15$, $p = 0.048$). This interaction was driven by an increase in accuracy for RRC-disambiguated sentences towards the end of the experiment for the Filler-exposed group that was absent for the RRC-exposed group (see Figure 6). We briefly interpret these results in the discussion section.

Discussion

In this experiment, we used the difference in RTs between RRC-disambiguated sentences and fillers at the disambiguating region as a proxy for GPE. The proxy-GPE decreased over time for the RRC-exposed group. This is consistent with the gradual decrease in GPE observed in previous studies and Experiment 1. If this decrease in GPE were driven by increased exposure to RRC-disambiguated sentences, we would expect the Filler-exposed group, which was not exposed to RRC-disambiguated sentences in the exposure phase, to show a stronger GPE in the test phase. Yet we failed to find a difference in the garden-path effects between the groups. This suggests that the decrease in garden-path effect for the RRC-exposed group was unlikely to be driven primarily by syntactic adaptation.

Earlier, we discussed the possibility that this decrease in garden-path effect over time could have been driven by asymmetric effects of task adaptation: sentences that were read more slowly when presented at the beginning of the experiment had a greater potential to be affected by task adaptation than sentences that were read more rapidly. We next describe exploratory analyses investigating this possibility.

Is the decrease in GPE a consequence of asymmetric task adaptation? Since people read sentences more rapidly as the experiment progresses (Experiment 1, Experiment 2, Fine and Jaeger 2016; Fine et al. 2013; Harrington Stack et al. 2018), a sentence is likely to be read more slowly if it is presented towards the beginning of the experiment than if it is presented towards the end. Let us define ΔRT_x for a given sentence x as this difference in the RT for x as a function of when in the experiment it is presented. The asymmetric task adaptation account predicts that ΔRT_x is proportional to the time taken to read x prior to task adaptation. If this is the case, then $\Delta RT_x > \Delta RT_y$ if the RT for x prior to task adaptation is greater than the RT for y prior to task adaptation.

We estimated the RT for a sentence prior to task adaptation by the time taken to read that sentence when it was presented at the beginning of the experiment. Using this estimate, we can rephrase the prediction of the asymmetric task adaptation account in the following manner: $\Delta RT_x > \Delta RT_y$ if x is read more slowly than y when both x and y are presented at the beginning of the experiment. We tested this prediction by comparing the RTs in Block 1 and Block 5 for sentences that were grouped by their mean RT in Block 1.

Since only a subset of our items occurred in both the blocks (at least once across the eight random-ordered lists), we use only this subset for the following analysis. We randomly split our participants into two halves. We then grouped sentences into quartiles based on their RTs in the disambiguating region in Block 1, averaged across participants in the first half. The first quartile consisted of the top 25% of the sentences that were read most rapidly in Block 1 by the first half of the participants. Similarly, the fourth quartile consists of the bottom 25% of sentences that were read most slowly in Block 1 by the first half of the participants. We then computed the mean RT for each quartile for the second half of the participants in Block 1 and Block 5 by averaging across the RT in the disambiguating region for all items in that quartile. We repeated this process for 1000 random splits of participants. In Figure 7 we plot the RT for the different quartiles in Block 1 and Block 5 averaged across the 1000 random splits of participants.

This analysis showed that sentences that were read more slowly when presented early in the experiment showed a greater ΔRT from Block 1 to Block 5 than sentences that were read more rapidly. This was true for both RRC-disambiguated sentences and fillers. These results support the hypothesis that the decrease in garden path effects observed in this experiment, and likely in previous experiments, may have been driven by asymmetric effects of task adaptation.

Of course, these results cannot rule out the possibility that there is an effect of syntactic adaptation over and above the task adaptation effect, but our experiment did not have the power to detect that effect. We ran a post-hoc power analysis to get an estimate of the number of participants that we would need to detect this additive effect if it exists.

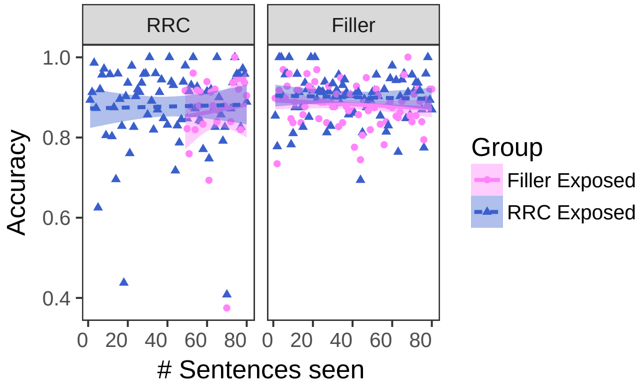


Figure 6. Accuracy over time in Experiment 2 with a best fit line estimated using a linear model.

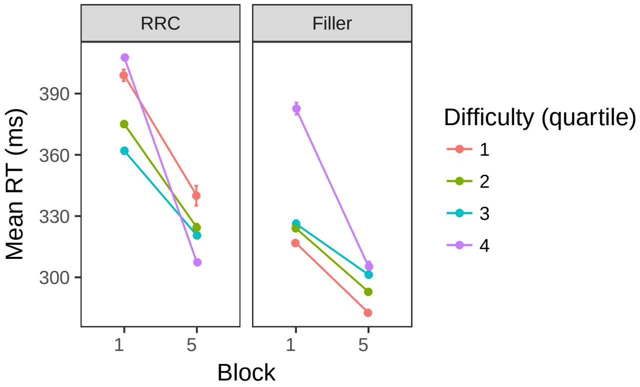


Figure 7. RTs in Block 1 and Block 5 for half the participants. Items are grouped into quartiles based on the RTs in Block 1 for the other half of the participants. The estimates are averaged across 1000 random splits of participants.

Post-hoc power analysis. For the power analysis we used a model that was almost identical to the one in Analysis 2, but with a simpler random effect structure (henceforth referred to as $Model_P$).[3] This included random intercepts for participant and item along with a by-participant slope for the interaction between sentence type and stimulus number and a by-item slope for group. In order to estimate the power to detect the additive effect, we ran 100 simulations using $Model_P$. For every simulation, we generated random orders of items and assigned a set of simulated participants to these orders. For every participant and item, we estimated the fixed effects using the coefficients of $Model_P$. For example, let us

[3] When we used the original random effect structure, most of the models in our simulation failed to converge.

Effect	# Participants	$\widehat{Power}\ p < 0.05$	$\widehat{Power}\ p < 0.01$
Main effect of group	200	1	0.98
	400	1	1
	800	1	1
Interaction between group and sentence type	200	0.17	0.07
	400	0.25	0.10
	800	0.33	0.14
Interaction between group, sentence type, and stimulus number	200	0.43	0.20
	400	0.69	0.46
	800	0.89	0.73

Table 1

Post-hoc power analyses. The estimate of power ($\widehat{Power}$) for any effect is defined as the proportion of simulations in which the effect was significant. Since we ran 100 simulations, when $\widehat{Power} = 1$, the true power is likely to be between 0.99 *and* 1

consider a simulated participant X in the RRC-exposed group (coded as 1) where the fifth item this participant was exposed to was a filler (coded as −1). The estimate of fixed effects for this participant for the fifth item was calculated using the following formula where g, *sn and st* stand for group, stimulus number and sentence type respectively, and $z(\log(x))$ returns the z-scored value of the natural log of any stimulus number x:

$$
\begin{aligned}
fixed_{X5} = \hat{\beta}_{intercept} &+ (1 \cdot \hat{\beta}_g) + (-1 \cdot \hat{\beta}_{st}) + (z(\log(5)) \cdot \hat{\beta}_{sn}) \\
&+ (1 \cdot -1 \cdot \hat{\beta}_{g:st}) + (1 \cdot z(\log(5)) \cdot \hat{\beta}_{g:sn}) \\
&+ (-1 \cdot z(\log(5)) \cdot \hat{\beta}_{st:sn}) + (1 \cdot -1 \cdot z(\log(5)) \cdot \hat{\beta}_{g:st:sn})
\end{aligned}
$$

We estimated the random intercepts and random slopes for every participant and item by randomly sampling from a distribution that was consistent with the covariance matrix of the random effects in $Model_P$. We added the fixed and random effects together, to get an estimate of the reading time for every participant and item combination. We then fit a new model (which had the same structure as $Model_P$) with these estimates.

Table 1 summarizes the proportion of simulations in which the following effects were significant at $p < 0.05$ and $p < 0.01$:

- Main effect of group: Overall difference in RTs for RRC-exposed and Filler-exposed participants.
- Interaction between group and sentence type: Difference in the proxy-GPE between RRC-exposed and Filler-exposed participants.
- Three-way interaction (group, sentence type and stimulus number): Difference between RRC-exposed and Filler-exposed participants in the rate of decrease in the proxy GPE over the course of the test phase.

With 800 participants (400 per group), a replication of Experiment 2 would have 89% power to detect the three-way interaction at $p < 0.05$ alpha level. Power is much lower with fewer participants or a more stringent alpha level threshold. The power to detect a two-way interaction between group and sentence type at $p < 0.05$ is only 33% even with as many as 800 participants. The fact that the three-way interaction is easier to detect than than two-way interaction suggests that even if there is a difference in the proxy-GPE between groups, this difference likely decreases over the course of the test phase.

We next consider the main effect of group: in Experiment 2, participants in the Filler-exposed groups read sentences more slowly in the test phase than participants in the RRC-exposed group. Our analysis suggests that this effect would almost certainly replicate, even with just 200 participants. This, taken together with the fact that this effect was observed in previous studies (Fine et al., 2013; Harrington Stack et al., 2018) suggests that this is a robust effect that is worth exploring in future studies.

Accuracy as a dependent measure of syntactic adaptation. In our models of question answering accuracy, there was a significant three-way interaction between sentence type, group and stimulus number: there was an increase in accuracy for RRC-disambiguated sentences in the test phase for participants in the Filler-exposed group, but not for participants in the RRC-exposed group (see Figure 6). This could be interpreted as evidence for rapid syntactic adaptation for the Filler-exposed group under the following assumption: the lower accuracy for RRC-disambiguated sentences when compared to fillers reflects the processing effort for RRC-disambiguated sentences — i.e. these sentences have lower accuracy because they are more difficult to process. If this is the case, then as participants adapt to RRC-disambiguated sentences, the accuracy on these sentences should increase. Therefore, under this account, the increase in accuracy for RRC-disambiguated sentences during the test phase can be interpreted as a consequence of the Filler-exposed group rapidly adapting to these sentences.

This account leaves several issues unexplained. First, why would sentences that are more difficult to process have lower accuracy, especially if participants do not need to correctly parse the sentence in order to answer the questions accurately? Second, why do we not observe an increase in accuracy for the RRC-exposed group at the beginning of the experiment when they first encounter these sentences? Resolving these questions is not within the scope of this paper and would need to be explored in future experiments.

General Discussion

Previous studies found that garden path effect for sentences with reduced relative clauses (like (1a) and (2a)) decreased over the course of the experiment Fine and Jaeger (2016); Fine et al. (2013). The authors interpreted this as evidence for syntactic adaptation. We presented an alternative account of the decrease in garden path effect: an asymmetric effect of task adaptation, whereby difficult sentences that are read more slowly when presented at the beginning of the experiment (e.g., (2a)) exhibit a larger task-driven decrease in reading times than those that are read more quickly (e.g., (2b)). The goal of this paper was to disambiguate between these two accounts.

The results of Experiment 1 and 2 taken together provide further evidence that garden path effects indeed decrease over time, suggesting that this is a robust effect. However, this decrease was independent of participants' prior linguistic experience in the experiment: we failed to find a difference in the garden-path effect between participants exposed to 24 RRC-disambiguated sentences and participants exposed to 24 fillers. Therefore, it is unlikely that this decrease in the garden path effect over time was driven by syntactic adaptation. Exploratory analyses revealed that the decrease in reading time for a sentence (RRC-disambiguated or filler) was proportional to the time taken to read the that sentence at the beginning of the experiment. This lends support to the hypothesis that the observed decrease in garden path effect over time was driven by asymmetric effects of task adaptation.

Our results do not rule out the possibility that syntactic adaptation could alter the reading times for RRC-disambiguated sentences as the experiment progresses, over and above the effects of task-adaptation. If such an additive effect exists, our experiment was likely not powered to detect it. We ran a post-hoc power analysis to determine how many participants a replication of Experiment 2 would need to detect the additive effect of syntactic adaptation if it exists. The analysis revealed that even if the additive effect did exist, it would be difficult to detect, even with a large number of participants. Furthermore, our power analysis measured only the ability to detect the *presence* of syntactic adaptation. Any manipulations designed to further probe the factors that affect syntactic adaptation—e.g., comparing groups exposed to different kinds of relative clauses—would likely result in an even smaller difference between groups, further decreasing power.

These results also have methodological implications for self-paced reading studies more generally, at least when run on crowd-sourced platforms like Amazon's Mechanical Turk. If sentences in the experimental condition exhibit a larger task-driven decrease in reading times than sentences in the control condition, then later trials are likely to underestimate the true difference between these conditions. In other words, the longer the experiment, the smaller our average estimate of the difference between these conditions. Therefore, in order to get better estimates of the true effect size, we recommend running shorter experiments with more participants.

Conclusion

Overall, we conclude that the decrease in garden path effects that has been consistently observed in self-paced reading studies may be driven entirely by task-adaptation; as such, it does not provide unambiguous evidence for syntactic adaptation. Furthermore, even if there were an additive effect of syntactic adaptation, we expect that such an effect will be difficult to detect, even with a large number of participants. Alternative paradigms are likely to be necessary to produce unambiguous evidence for syntactic adaptation and reliably study the factors that can modulate it.

Acknowledgments

We would like to thank Nicholas Douglass and Brian Leonard for assistance with stimulus creation and data collection respectively. We would also like to thank Brian Dillon and the members of the Computational and Psycholinguistics lab at Johns Hopkins University for their helpful comments.

References

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, Inc.

Drummond, A., von der Malsburg, T., Erlewine, M. Y., & Vafaie, M. (2016). *Ibex farm.* https://github.com/addrummond/ibex. GitHub.

Fine, A. B., & Jaeger, T. F. (2016). The role of verb repetition in cumulative structural priming in comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(9), 1362–1376. Retrieved from http://dx.doi.org/10.1037/xlm0000236

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid Expectation Adaptation during Syntactic Comprehension. *PLoS One*, *8*(10), e77661. Retrieved from https://doi.org/10.1371/journal.pone.0077661

Harrington Stack, C. M., James, A. N., & Watson, D. G. (2018). A failure to replicate rapid syntactic adaptation in comprehension. *Memory and Cognition*, *46*(6). doi: 10.3758/s13421-018-0808-6

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26.

MacDonald, M. C., Just, M. A., & Carpenter, P. A. (1992). Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology*, *24*(1), 56–98. Retrieved from https://doi.org/10.1016/0010-0285(92)90003-K

MacDonald, M. C., Pearlmutter, N. J., & Seidenberg, M. S. (1994). The lexical nature of syntactic ambiguity resolution. *Psychological Review*, *101*(4), 676–703. Retrieved from http://dx.doi.org/10.1037/0033-295X.101.4.676

Malone, A., & Mauner, G. (2018). What do readers adapt to in syntactic adaptation? In *Poster session presented at the 31st Annunal CUNY Sentence Processing Conference.*

Mitchell, D. C., Cuetos, F., Corley, M. M. B., & Brysbaert, M. (1995). Exposure-Based Models of Human Parsing: Evidence for the Use of Coarse-Grained (Nonlexical) Statistical Records . *Journal of Psycholinguistic Research*, *24*(6), 469–488.

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914.

Trueswell, J. C. (1996). The Role of Lexical Frequency in Syntactic Ambiguity Resolution. *Journal of Memory and Language*, *35*(4), 566–585. Retrieved from https://doi.org/10.1006/jmla.1996.0030

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive Psychology*, *58*, 250–271.