Eye-tracking evidence for active gap-filling regardless of dependency length

Wing-Yee Chow[1] & Yangzi Zhou[1,2]

[1] Division of Psychology and Language Sciences, University College London, UK

[2] Department of Psychology, University of Edinburgh, United Kingdom

Corresponding author:

Wing-Yee Chow (wingyee.chow@ucl.ac.uk)

Chandler House, 2 Wakefield Street, London, WC1N 1PF, United Kingdom

Abstract

Previous work on real-time sentence processing has established that comprehenders build and interpret filler-gap dependencies without waiting for unambiguous evidence about the actual location of the gap ("*active gap-filling*") as long as such dependencies are grammatically licensed. However, this generalisation was called into question by recent findings in a self-paced reading experiment by Wagers and Phillips (2014; W&P14) which may be taken to show that comprehenders do not interpret the filler at the posited gap when the dependency spans a longer distance. In the present study we aimed to replicate these findings in an eye-tracking experiment with better controlled materials and increased statistical power. Crucially, we found clear evidence for active gap-filling across all levels of dependency length. This diverges from W&P14's findings but is in line with the long-standing generalisation that comprehenders build and interpret filler-gap dependencies predictively as long as they are grammatically licensed. We found that the effect became smaller in the long dependency conditions in the post-critical region, which suggests the weaker effect in the long dependency conditions may have been undetected in W&P's study due to insufficient statistical power and/or the use of a self-paced reading paradigm.

Keywords: language comprehension; sentence processing; eye-tracking; filler-gap dependencies; prediction

Introduction

Natural language is full of dependencies between non-adjacent elements. For example, in a simple wh-question like (1), the wh- phrase *which book* appears at the beginning of the question but it is interpreted as the direct object of the verb *read* which appears at the end of the question. Crucially, as (2) illustrates, the displaced element (a *filler*) can be indefinitely far away from the position at which it gets interpreted (its *gap).*

(1) Which book did John read ____?

(2) Which book did Bill say that Mary thought John read ___?

Psycholinguists have long been interested in how comprehenders compute such unbounded dependencies between a filler and its gap (also known as filler-gap dependencies) in real time. Results from studies using different experimental techniques have commonly suggested that such dependencies are computed predictively, that is, without waiting for unambiguous evidence about the actual location of the gap ("*active gap-filling*", e.g., Crain & Fodor 1985; Frazier, 1987; Frazier & Flores d'Arcais, 1989; Garnsey, Tanenhaus, & Chapman, 1989; Lee 2004; Omaki et al., 2015; Pickering & Traxler, 2001; Stowe, 1986; Sussman & Sedivy, 2003; Traxler & Pickering, 1996; Wagers & Phillips, 2009).

One type of evidence for active gap-filling comes from comprehenders' response to information that indicates a potential gap has been filled ("filled-gap" effect; Crain & Fodor, 1985; Lee, 2004; Stowe, 1986). For example, Stowe (1986) asked whether comprehenders posit a gap for a filler actively by examining their reading times in sentences such as (3). In (3a), the verb *bring* constitutes the first possible gap site for the filler *who*, but the presence of the direct object *us* indicates that the filler cannot be interpreted there and a gap must be identified elsewhere; in (3b) there is no filler and as such no gap-filling is required.

(3) Filled-gap paradigm (Stowe, 1986):

a. My brother wanted to know who Ruth will bring us home to at Christmas.

b. My brother wanted to know if Ruth will bring us home to Mom at Christmas.

Results revealed that comprehenders showed longer reading times upon encountering the direct object *us* in (3a) compared to (3b). This suggests that comprehenders posit a gap for the filler at the earliest possible position (i.e., at the verb *bring*) and are surprised when subsequent information (the direct object) indicates that the predicted gap has already been filled and is therefore not viable.

Another type of evidence for active gap-filling concerns comprehenders' immediate sensitivity to the plausibility of the thematic relationship between the filler and its potential gap (plausibility effect; e.g. Traxler & Pickering 1996; Garnsey et al., 1989). For instance, Garnsey et al. (1989) examined comprehenders' event-related brain potentials (ERP) response as they read sentences with a semantically plausible (4a) or implausible (4b) filler-gap dependency.

(4) Plausibility paradigm (Garnsey et al., 1989):

a. The businessman knew which customer the secretary called at home.

b. The businessman knew which article the secretary called at home.

Comprehenders showed a larger centro-posterior negativity peaking at around 400ms after the onset of the target verb (also known as an N400 effect) when the filler was an implausible direct object (4b) than when it was a plausible direct object (4a). Such immediate sensitivity has been taken to show that comprehenders actively interpret a filler at the first possible gap and are therefore surprised when it results in a semantically implausible interpretation. In addition, studies that examined comprehenders' reading times in sentences like these have also shown that reading times are longer when the dependency between a filler and its first possible gap is semantically implausible (Omaki et al., 2015; Traxler & Pickering, 1996).

Taken together, these findings suggest that “active gap-filling” not only involves the formation of a syntactic dependency between the filler and a gap (which can be studied using a filled-gap paradigm), but also the semantic interpretation of the filler at the gap (which can be studied using a plausibility paradigm).

Constraints on Active Gap-filling

Further, previous work has demonstrated that this predictive processing mechanism is grammatically constrained, such that active gap-filling occurs only in grammatically licensed environments (Omaki et al., 2015; Phillips, 2006; Stowe, 1996; Traxler & Pickering, 1996; Wagers & Phillips, 2009). A number of studies have shown that comprehenders do not posit gaps inside syntactic domains which block filler-gap dependencies (also known as “syntactic island” constraints; Ross, 1967). For example, while a dependency is possible between the filler *which movie* and the verb *seen* in (5), such a dependency is not possible when the verb *seen* is embedded within a relative clause (6).

(5) I wonder which movie the student had seen __.

(6) I wonder which movie the student who had seen (*__) Dunkirk was talking about __.

Traxler and Pickering (1996) asked whether evidence for active gap-filling can be observed in a syntactic island with sentences like (7) and (8). They manipulated the plausibility of the filler as the direct object of the target verb (e.g., which book/city … wrote) and the syntactic position of the target verb (inside a relative clause island or not). They examined whether comprehenders’ sensitivity to the plausibility of the filler as the direct object of the target verb is modulated when the verb is inside a syntactic island.

(7) Non-island conditions:

a. *Plausible:* We like the book that the author <u>wrote</u> unceasingly and with great dedication about while waiting for a contract.

b. *Implausible:* We like the city that the author wrote unceasingly and with great dedication about while waiting for a contract.

(8) Island conditions:

a. *Plausible:* We like the book that the author who wrote unceasingly and with great dedication saw while waiting for a contract.

b. *Implausible:* We like the city that the author who wrote unceasingly and with great dedication saw while waiting for a contract.

If comprehenders posit a gap for a filler only in grammatically licensed syntactic environments, then they should not attempt to interpret the filler as the direct object of the target verb when the verb is inside a relative clause island and therefore should not be sensitive to the plausibility manipulation in (8). This prediction was confirmed by the results, which showed a slowdown in the implausible condition compared to the plausible condition in the non-island sentences (7) but no effect of plausibility in the island sentences (8). These findings suggested that active gap-filling is not always at work and instead it is constrained by grammatical knowledge.

However, recent findings reported by Wagers and Phillips (2014; henceforth W&P14) suggested that active gap-filling may be absent even when the dependency is fully licensed by grammatical knowledge. The authors asked whether active gap-filling may be modulated by the distance between a filler and its potential gap using the filled-gap and plausibility paradigms in two self-paced reading experiments. They manipulated the linear distance between the filler and its potential gap by adding a prepositional phrase (PP) between the filler and the target verb, and the structural (as well as linear) distance by introducing an additional clause (CP) to embed the target verb more deeply.

Crucially, even though they observed a filled-gap effect in the target region across all three levels of dependency length in an initial experiment, they found a plausibility effect in the target region (which consisted of the three words immediately following the critical verb) only in the short dependency condition in a subsequent experiment. In both the long dependency conditions (PP and CP), comprehenders showed a plausibility effect only after the actual gap became evident. In other words, even though comprehenders' sensitivity to a filled gap was not modulated by the distance between the filler and gap, when the filler and gap were linearly and/or structurally further apart comprehenders did not show any sensitivity to the plausibility of the dependency before reaching the actual gap.

Based on the standard interpretation of the filled-gap and plausibility effects, these results may be taken to show that comprehenders always form a syntactic dependency between a filler and a gap predictively, but they may not interpret the filler at the posited gap when the dependency spans a longer distance. In other words, active gap-filling may not be fully at work when dependency is long even if it is grammatically licensed. Under this view, these results constitute a clear exception to the generalisation that comprehenders build and interpret filler-gap dependencies predictively as long as they are grammatically licensed.

However, this interpretation was not adopted by W&P14. Instead, they kept the active gap-filling generalisation and took the differential effects of dependency length on comprehenders' sensitivity in the two experiments to suggest that certain types of memory representations may be privileged relative to others. They argued that comprehenders can detect a filled gap as long as they maintain the grammatical category of the filler in working memory while looking for a gap for the filler, but in order to evaluate the plausibility of a potential filler-gap dependency comprehenders must also maintain the lexically specific semantic features of the filler in memory. Specifically, W&P14 took their findings to propose that comprehenders can maintain representations of the filler's grammatical category, but not its semantic features, in working memory over a long distance for active gap-filling, and that comprehenders must retrieve the filler's semantic features from memory at the actual gap in order to evaluate the plausibility of the dependency.

This proposal not only posits a distinction between two types of memory representations (representations of syntactic category and semantic features respectively) required for computing filler-gap dependencies, but crucially it also posits a distinction between how easily these two types of representations may be maintained in memory or how quickly they decay over time. As such, it could have important theoretical implications, both for theories of sentence processing (e.g., Lewis, Vasishth & Van Dyke, 2006; Ness & Meltzer-Asscher, 2017; Santi, Friederici, Makuuchi & Grodzinsky, 2015) as well as models of memory representations (e.g., Baddeley, 2010; Tulving, 1972).

The Present Study

A key piece of evidence from W&P14's study which was unexpected and which motivated their proposal was the absence of a plausibility effect in the long dependency conditions in Experiment 2. Therefore, in the present study we aim to replicate the results of W&P14's Experiment 2 and examine whether a plausibility effect may be observed in the long dependency conditions. In order to provide a stronger test for the effect of dependency length on comprehenders' sensitivity to a potential dependency's plausibility, we introduced two main changes to the methods and materials.

First, whereas W&P14 used a self-paced reading paradigm and recorded comprehenders' response time to each word in a sentence, we used an eye-tracking paradigm and recorded comprehenders' eye movements as they read sentences freely on a computer screen. In an eye-tracking paradigm, sentences are presented as a whole on a computer screen and comprehenders are free to read and reread parts of the sentences as they wish. Meanwhile, in a self-paced reading paradigm, comprehenders read sentences one word at a time by pressing a button to proceed in a strictly left-to-right fashion and they are not allowed to reread earlier parts of the sentence. Arguably, this places higher cognitive demands on comprehenders as they have to make repeated button presses while reading and comprehend sometimes long and complex sentences without being able to reread parts of the sentences. This may also explain why reading times in a self-paced reading paradigm tend to be longer than normal reading (Jegerski, 2014) and that effects are more likely to be delayed or “spill over” into later regions in self-paced reading than in eye-tracking (Witzel, Witzel & Forster, 2012). Therefore, in the present study we used an eye-tracking paradigm to examine the processing of long-distance dependencies in a more naturalistic setting.

In addition, W&P14 only used 24 sets of items, resulting in only 4 trials per participant per condition. This is considerably lower than that typically found in other self-paced reading studies (e.g., Omaki et al., 2015) and may have contributed to a lack of statistical power for detecting potentially small differences. Therefore, we doubled the number of experimental items in order to increase the statistical power of the present study.[1]

Methods

Participants

Thirty young adults (20 female, 10 male, mean age = 23 years, aged between 20 and 32) participated in the present study. All participants were native speakers of British English and had normal or corrected-to-normal vision. All participants gave informed consent and received 7.5 GBP for their participation. Data from three additional participants were excluded due to low comprehension accuracy (< 80%). For the remaining 30 participants, the mean comprehension accuracy was 88%.

[1] W&P14 tested 36 participants on 24 items while the present study tested 30 participants on 48 items. As a result the present study (1440 trials) had 1.7x as many trials as W&P (864 trials).

Design and Materials

A total of 48 sets of experimental sentences were used in the present study. A sample item set is shown in Table 1. Twenty-four sets of sentences were taken from W&P14's Experiment 2 and were modified for British English spelling only. Another 24 sets of sentences were newly created and they were modelled on W&P14's original materials (see below). The full set of experimental stimuli can be accessed from https://osf.io/qctk4/.

Following W&P14's Experiment 2, we manipulated plausibility (plausible vs. implausible) and dependency length (Short vs. +PP vs. +CP) in a 2 × 3 within-participants design. Plausibility was manipulated by using filler NPs that could or could not be a semantically plausible argument of the target verb. In the Short conditions, the filler ("the posters/smiles") and the critical verb ("plastered") were separated by a two-word subject noun phrase and an adverb. In the +PP conditions, the subject noun phrase was modified by a five-word prepositional phrase (e.g., "with a large cloth bag"). Finally, in the +CP conditions, the relative clause in the short conditions was further embedded in another five-word clause (e.g., "the energetic teenager said that …"). Finally, a five-word preamble (e.g., "It excited the teenager that") was added to the beginning of the sentences in the Short conditions in order to keep the position of the critical word identical across conditions.

Table 1. Sample stimuli in the current study.

Dependency length	***Plausibility***	***Sentence***
Short	Plausible	It excited the teenager that the posters which the campaigner tirelessly plastered the big bulletin board with in the city centre were designed by an artist.
	Implausible	It excited the teenager that the smiles which the campaigner tirelessly plastered the big bulletin board with in the city centre were designed by an artist.
+PP	Plausible	The posters which the campaigner with a large cloth bag tirelessly plastered the big bulletin board with in the city centre were designed by an artist.
	Implausible	The smiles which the campaigner with a large cloth bag tirelessly plastered the big bulletin board with in the city centre were designed by an artist.
+CP	Plausible	The posters which the energetic teenager said that the campaigner tirelessly plastered the big bulletin board with in the city centre were designed by an artist.
	Implausible	The smiles which the energetic teenager said that the campaigner tirelessly plastered the big bulletin board with in the city centre were designed by an artist.

Following W&P14, all critical verbs in the current study were spray-load-type alternating locative verbs (Fraser, 1971; Rappaport & Levin, 1986). These verbs take two internal arguments, figure and ground, which can alternate between two configurations (e.g., ground-figure: "The campaigner plastered the bulletin board with the poster" vs. figure-ground: "The campaigner plastered the poster onto the bulletin board"). Following W&P14, only the ground–figure configuration was used in the experimental materials, and implausible filler NPs were implausible as either figure or ground.

For the 24 newly created items, we selected 12 verbs that were not used in W&P14's original materials based on Levin's (1993) classification and used each as the critical verb in two item sets. The filler NPs in these newly created items also did not overlap with those used in W&P14. In addition, we used inanimate filler NPs in all of the new materials and matched the plausible and implausible filler NPs on length, frequency (Van Heuven, Mandera, Keuleers & Brysbaert, 2014) and concreteness (Brysbaert, Warriner & Kuperman, 2014). Further, we extended the sentences beyond the critical verb with post-target continuations that were identical across conditions, and ensured that the post-target continuations did not introduce any other filler-gap dependency.

The 48 item sets were divided into 6 presentation lists, such that each list contained exactly one version of each item. Each list also contained 48 filler sentences, which varied in length and syntactic complexity, and some of which also contained a filler-gap dependency (e.g., subject-extracted relative clause). Each sentence was followed by a Yes/No comprehension question to ensure that participants were attending to the stimuli. The order of experimental and filler sentences was randomised across participants.

Procedure

Participants were tested individually in a quiet room. A desktop mount EyeLink 1000 eye tracker (SR Research, Toronto, Ontario, Canada) was used to monitor participants' eye movements. Materials were displayed on a LCD monitor which was placed 70 cm from the participants. At this distance, 3.7 characters were displayed per degree of visual arc. The eye-tracker has an angular resolution of 0.25 – 0.5 degrees. All experimental sentences were displayed on two lines with a line break placed after the preposition "with" which unambiguously marked the gap position. Data were recorded at a sampling rate of 1000 Hz.

A calibration procedure was performed before the experiment began, and re-calibration was performed as needed throughout the experiment. At the start of each trial, a black square appeared on the left central portion of the screen, marking the position of the first character of text. The stimulus text was presented as soon as a fixation was detected on the square. Participants were instructed to read each sentence at a natural pace, and to press a button when they had finished reading. A Yes/No comprehension question was presented after each trial and participants were instructed to respond by pressing one of two buttons on a hand-held controller. Prior to the experimental session, participants were presented with three practice trials to familiarise themselves with the task. An average experimental session lasted about 60 minutes.

Analysis

Before analysis, fixations of less than 80 ms in duration and within one character of the previous or following fixation were incorporated into this neighbouring fixation, and any remaining fixations shorter than 80 ms were excluded. Data from three item sets were excluded from data analysis due to experimental programming error. Further, in the remaining data, trials were excluded if there was a blink or track loss during first pass reading of the critical region (see below for region definitions). This resulted in deletion of 6.9% of trials.

Experimental sentences were divided into three regions of interest for data analysis: the adverb preceding the critical verb (pre-critical region), the critical verb itself (critical region), and the object NP immediately following the critical verb (post-critical region).

(9) The posters which the campaigner with a large cloth bag/ tirelessly/ <u>plastered/</u> the bulletin board/ with in the city centre were designed by an artist.

Three eye-tracking measures were computed and reported for each region. *First pass time* is the aggregate of all fixations in the region before the eye leaves the region for the first time, either to the left or to the right. *Regression path time,* also known as go-past time, is the sum of all fixations in a region before the eye leaves the region to the right, which includes any regressive eye fixations in previous regions. *Total time* is sum of all fixations in a region. Since total time may include fixations that participants made after having encountered the ultimate gap site (after the preposition “with”), we restrict our conclusions about active gap-filling to eye-tracking measures which only include fixations that occur unambiguously prior to the actual gap (i.e. first pass time and regression path time). For first pass and regression path time, a trial is excluded if the region was skipped on first pass reading; a trial is also excluded for total time if the region was not fixated at all.

Eye-movement data were analysed using linear mixed-effects models (LMMs), incorporating Length and Plausibility and their interaction as fixed effects, and subjects and items as random effects (see Baayen, Davidson, & Bates, 2008). One difference between our models and W&P14's is that we used the maximal random effects structure[2] (Barr, Levy, Scheepers & Tily, 2013), whereas W&P14 only included by-subject and by-item random intercepts.[3] Statistical analyses were performed on data from all items. Subsequently we also divided the data into two subsets (those from the newly created items and those from W&P14's items) to see if they showed qualitatively similar patterns of results. Note, however, that statistical analyses were not performed on these subsets of data as they are likely to be underpowered and therefore not informative.

Following W&P14, we coded the two levels of Plausibility with sum contrasts and the three levels of Dependency Length with Helmert contrasts: Short conditions were compared to the mean of +PP and +CP conditions (reported as the length.long coefficient); +PP conditions were compared to +CP conditions (reported as the length.clause coefficient). Factor labels were transformed into numerical values, and centred prior to analysis to minimise collinearity between variables (Baayen, 2008). We used the lmer function from the lme4 R package to fit the LMMs (version 1.1-12; Bates, Maechler, Bolker & Walker, 2015). The analysis yields regression coefficients (β), which estimate the effect size in milliseconds, and the t-value of the effect coefficient. A given coefficient was judged to be significant at $\alpha = 0.05$ if the absolute value of t exceeded 2 (Baayen et al., 2008).

[2] In all cases the maximal model that converged had zero correlation parameters in the random effects structure.

[3] We also analysed the present results with models with random intercepts only and found the same pattern of results.

All main effects and interactions involving either factor are reported in the results tables. However, we discuss effects of dependency length only when they interact with the effects of plausibility. This is because (i) the present experiment was not designed to examine the main effects of dependency length, and (ii) a prepositional phrase appeared prior to the pre-critical region in the +PP conditions and not in the Short and +CP conditions and therefore main effects of dependency length are not meaningful.

Further, in order to directly examine the effect of plausibility on each level of dependency length, we used the R package *bootES* (Gerlanc & Kirby, 2012; Kirby & Gerlanc, 2013) to compute effect size estimates in Cohen's *d* and their boostrap confidence intervals (CIs) for the contrast between the plausible and implausible conditions. CIs were computed using the bias-corrected-and-accelerated (BCa) boostrap method with 2000 resamples.

Results

Primary data analysis - All experimental materials

Condition means and the effect of plausibility (difference between the plausible and implausible conditions) on each level of dependency length are presented in Table 2. Standardised effect size (Cohen's *d*) and 95% confidence interval for the plausibility contrasts are presented in

Table 3. Results of the linear mixed effect models are presented in Table 4. Average regression path times in the critical and post-critical regions are shown in Figure 1.

In the pre-critical (adverb) region, there were no effects of plausibility in first pass time and regression path time. Linear mixed effect models revealed a main effect of plausibility in total time only, which shows that participants ultimately spent more time reading this region when the sentence was implausible. Plausibility did not interact with dependency length in any of the eye-tracking measures.

In the critical (verb) region, linear mixed models revealed a significant main effect of plausibility in both regression path time and total time, showing that reading times were longer in the implausible conditions than in the plausible conditions. The same pattern was also observed in first pass time, although the effect failed to reach statistical significance. Crucially, as shown in

Table 3, implausibility lead to longer regression path time and total time across all three levels of dependency length. This pattern of results contrasts with W&P14, in which a significant plausibility effect was found in the Short condition only.

In the post-critical region, there were no significant effects in first pass time. However, there was a significant effect of plausibility along with a plausibility × length.long interaction in both regression path time and total reading time. This indicates that the effect of plausibility, which was observed in the critical region across all three length conditions, weakened in the post-critical region in the +CP and +PP conditions compared to the Short conditions.

Table 2. Means (and standard errors) along with the difference the implausible and plausible conditions for first pass, regression path and total times (in milliseconds).

	Pre-critical region			Critical region (verb)			Post-critical region		
	Plausible	*Implausible*	Δ	*Plausible*	*Implausible*	Δ	*Plausible*	*Implausible*	Δ
First pass time									
Short	313 (16)	306 (13)	-7	284 (11)	288 (12)	5	714 (48)	744 (37)	30
+PP	334 (17)	342 (19)	8	291 (14)	305 (15)	15	761 (50)	734 (39)	-27
+CP	304 (17)	315 (14)	12	290 (16)	297 (14)	7	751 (44)	702 (43)	-49
Regression path time									
Short	426 (48)	452 (28)	27	369 (27)	442 (47)	73	984 (73)	1322 (99)	338
+PP	511 (63)	432 (40)	-79	391 (26)	473 (47)	82	1028 (61)	1116 (59)	88
+CP	500 (75)	437 (33)	-63	363 (27)	445 (44)	82	1226 (129)	1216 (115)	-10
Total time									
Short	514 (44)	650 (44)	136	435 (29)	568 (44)	133	1065 (70)	1502 (136)	437
+PP	549 (37)	680 (57)	131	477 (34)	589 (48)	112	1242 (81)	1437 (112)	195
+CP	567 (48)	621 (43)	54	490 (43)	581 (45)	90	1260 (109)	1376 (101)	116

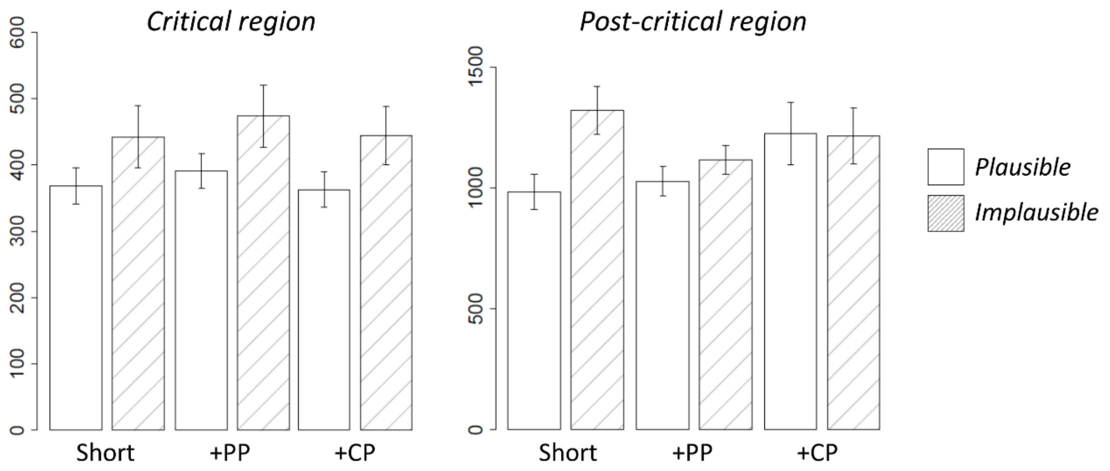

Figure 1. Average regression path time (ms) and standard error in the critical and post-critical regions.

Table 3. Standardised effect size (Cohen's d) and 95% confidence interval for the contrasts between the plausible and implausible conditions.

	Pre-critical region		Critical region (verb)		Post-critical region	
	Cohen's *d*	95% CI	Cohen's *d*	95% CI	Cohen's *d*	95% CI
First pass time						
Short	-0.08	[-0.43, 0.28]	0.08	[-0.30, 0.44]	0.15	[-0.20, 0.54]
+PP	0.08	[-0.32, 0.40]	0.20	[-0.14, 0.60]	-0.13	[-0.52, 0.23]
+CP	0.17	[-0.23, 0.55]	0.09	[-0.31, 0.46]	-0.20	[-0.53, 0.18]
Regression path time						
Short	0.10	[-0.30, 0.56]	0.27	[-0.08, 0.58]	0.77	[0.44, 1.06]
+PP	-0.22	[-0.49, 0.15]	0.41	[0.03, 0.70]	0.25	[-0.16, 0.59]
+CP	-0.17	[-0.40, 0.29]	0.36	[0.04, 0.56]	-0.01	[-0.36, 0.38]
Total time						
Short	0.53	[0.12, 0.86]	0.72	[0.35, 1.07]	0.78	[0.48, 1.15]
+PP	0.56	[0.23, 0.88]	0.60	[0.19, 0.92]	0.40	[0.10, 0.64]
+CP	0.32	[-0.06, 0.71]	0.54	[0.09, 0.94]	0.33	[-0.09, 0.68]

Table 4. Linear mixed effects model results (coefficients, standard errors and *t*-values) for first pass, regression path and total times.

	Pre-critical region			Critical region (verb)			Post-critical region		
	β	*SE*	*t*	*β*	*SE*	*t*	*β*	*SE*	*t*
First pass time									
(Intercept)	319	13	24.61*	291	11	27.33*	739	39	19.09*
plausibility	-6	9	-0.65	-9	7	-1.32	19	21	0.93
length.long	15	9	1.68	10	8	1.30	5	22	0.21
length.clause	30	12	2.45*	6	9	0.66	17	24	0.72
plausibility:length.long	-14	18	-0.78	-6	16	-0.35	63	41	1.54
plausibility:length.clause	4	23	0.19	-5	18	-0.28	-24	61	-0.40
Regression path time									
(Intercept)	454	30	15.12*	411	27	15.32*	1149	72	16.02*
plausibility	25	28	0.89	-71	30	-2.36*	-140	58	-2.42*
length.long	24	30	0.82	20	27	0.76	-9	50	-0.19
length.clause	14	51	0.27	28	31	0.92	-146	81	-1.82
plausibility:length.long	78	61	1.27	-26	53	-0.50	311	115	2.70*
plausibility:length.clause	35	66	0.53	-2	62	-0.04	-95	155	-0.61
Total time									
(Intercept)	596	39	15.25*	521	36	14.68*	1319	92	14.31*
plausibility	-110	20	-5.37*	-114	24	-4.80*	-252	50	-5.00*
length.long	23	25	0.91	36	21	1.73	45	39	1.16
length.clause	23	24	0.96	0	24	0.02	15	50	0.29
plausibility:length.long	42	57	0.73	34	37	0.91	282	110	2.57*
plausibility:length.clause	-80	57	-1.39	-23	46	-0.49	-85	126	-0.67

Data from W&P14's original items vs. the newly created items

Condition means and the effect of plausibility on each level of dependency length for W&P14's original items and the newly created items are shown in Tables 5 and 6 respectively.[4] We observed qualitatively similar results in both sets of data; implausibility led to numerically longer regression path time in the critical and/or post-critical regions across all three levels of dependency length in both data sets. This suggests that comprehenders behaved largely similarly across the two sets of items.

[4] Tables 5 and 6 show averages of data from 29 participants. This is because one participant had missing values in one of the cells when data from the two subsets of items were analysed separately. Data from this participant were excluded from the calculation of grand averages in all regions and all measures in both subsets.

Table 5. Means (and standard errors) along with the difference the implausible and plausible conditions for first pass, regression path and total times (in milliseconds) in W&P14's original items.

	Pre-critical region			Critical region (verb)			Post-critical region		
	Plausible	*Implausible*	Δ	*Plausible*	*Implausible*	Δ	*Plausible*	*Implausible*	Δ
First pass time									
Short	332 (23)	298 (13)	-35	283 (14)	288 (14)	4	768 (55)	771 (51)	3
+PP	348 (21)	367 (25)	20	279 (15)	293 (21)	14	777 (58)	724 (41)	-53
+CP	282 (14)	318 (18)	36	275 (16)	317 (18)	42	748 (49)	739 (54)	-9
Regression path time									
Short	435 (41)	433 (27)	-2	324 (19)	603 (142)	279	1096 (108)	1383 (170)	287
+PP	570 (97)	424 (40)	-146	343 (28)	533 (85)	190	987 (69)	1044 (70)	56
+CP	441 (65)	395 (50)	-45	370 (36)	433 (49)	63	1342 (171)	1336 (185)	-6
Total time									
Short	549 (52)	649 (45)	100	426 (33)	575 (48)	149	1165 (88)	1525 (137)	360
+PP	538 (51)	692 (53)	154	438 (39)	552 (44)	114	1228 (111)	1350 (105)	121
+CP	568 (50)	575 (50)	7	494 (45)	582 (55)	89	1265 (114)	1354 (110)	89

Table 6. Means (and standard errors) along with the difference the implausible and plausible conditions for first pass, regression path and total times (in milliseconds) in the newly created items.

	Pre-critical region			Critical region (verb)			Post-critical region		
	Plausible	*Implausible*	Δ	*Plausible*	*Implausible*	Δ	*Plausible*	*Implausible*	Δ
First pass time									
Short	310 (24)	326 (19)	17	286 (14)	287 (17)	1	687 (45)	741 (41)	54
+PP	338 (22)	326 (17)	-12	305 (20)	314 (17)	9	768 (54)	771 (52)	2
+CP	331 (22)	321 (14)	-10	306 (20)	289 (17)	-17	779 (54)	694 (41)	-85
Regression path time									
Short	457 (108)	458 (47)	.55	419 (57)	408 (46)	-11	886 (54)	1302 (111)	415
+PP	444 (43)	443 (63)	-.25	440 (44)	421 (38)	-19	1070 (86)	1224 (76)	154
+CP	588 (127)	468 (39)	-120	360 (31)	468 (62)	108	1156 (122)	1176 (95)	20
Total time									
Short	489 (45)	662 (57)	173	444 (30)	559 (53)	115	1007 (63)	1523 (155)	517
+PP	575 (40)	691 (74)	116	517 (38)	634 (69)	116	1290 (79)	1572 (150)	283
+CP	584 (56)	685 (47)	101	493 (50)	594 (46)	101	1297 (117)	1451 (112)	154

Discussion

The present study has two main findings. First, we found a clear effect of plausibility on comprehenders' regression path and total reading times in the critical region across all three levels of dependency length. Further, we found that plausibility interacted with dependency length in the post-critical region, which indicates that the effect of plausibility was less sustained when the dependency spans a longer distance.

Active gap-filling regardless of dependency length

The most important finding in the present study is the observation of a plausibility effect across all three levels of dependency length in the critical region. We take this finding to show that comprehenders engage in active gap-filling even when the potential gap is far away from the filler, and that they can immediately detect any implausibility because the semantic features of the filler are maintained in memory for active gap-filling. This argues against W&P14's proposal that comprehenders cannot maintain the semantic features of a filler over a long distance for active gap-filling, and instead must retrieve them from memory upon encountering the actual gap. Further, since the dependency between the filler and the gap is grammatically licensed in all three length conditions, our finding is in line with the generalisation that comprehenders build and interpret filler-gap dependencies predictively as long as they are grammatically licensed.

W&P14's proposal was based on the observation that implausibility led to longer reading times in the active filling region in the Short conditions, but not in the Long (+PP and +CP) conditions. We propose two possible explanations for the apparent lack of a plausibility effect in the Long conditions in W&P14's study:

1. *Insufficient statistical power.* One key difference between W&P14 and the present study is that W&P14 used 24 sets of experimental items while we had 48 sets. With a total of 6 conditions, each participant in W&P14 saw at most 4 trials in each condition. This is quite low compared to other self-paced reading studies in the literature. For instance, both Lee (2004) and Omaki et al. (2015) used 28 sets of items in a 4-condition experiment, resulting in 7 trials per condition. Further, as shown in Table 2, in the present study the plausibility effect which spilled over into the post-critical region was weaker in the Long conditions than in the Short conditions. Therefore, the effect of plausibility may also have been weaker and thus more difficult to detect in the Long conditions in W&P14.

2. *The lack of rereading in the self-paced reading paradigm.* Another important difference between these studies is the use of self-paced reading in W&P14 and eye-tracking in the present study. In a self-paced reading paradigm, participants read sentences word by word in a strictly left-to-right fashion and are not allowed to reread earlier parts of a sentence. Therefore, this method may not be well-suited to capture effects that would otherwise arise in re-reading. In fact, in the present study the clearest effect of plausibility was observed in regression path and total reading times, both of which are modulated by re-reading behaviour. Further, as shown in Table 2, the effect of plausibility on first pass time, an eye-tracking measure which is arguably analogous to reading times in a self-paced reading paradigm, was very small across all conditions in the critical region, and it even went in the opposite direction in the Long conditions in the post-critical region.[5] This suggests that in the present study participants' sensitivity to the plausibility manipulation became apparent only when they had had the opportunity to reread earlier parts of the sentence, which may explain why W&P14 did not detect an effect of plausibility in the Long conditions with self-paced reading.

In sum, the present results showed that comprehenders are sensitive to the semantic features of the filler prior to the actual gap even when the potential gap is far away from the filler. We propose that this is because comprehenders maintain information about the filler's semantic features in memory for active gap-filling (the maintenance view). Note, however, it is also possible that information about the filler is actively reactivated (e.g., Nicol & Swinney, 1989) or retrieved from memory (e.g., McElree & Griffith, 1998) at potential gap sites (the retrieval view, or "active filler-retrieval"). We believe these two processes need not

[5] One may argue that regression path time is also analogous to self-paced reading time as they both measure the total time spent reading before progressing to the next region of text. However, since the option to re-read earlier parts of the sentence is only available to comprehenders in an eye-tracking paradigm but not in self-paced reading, any suggestions that link self-paced reading time to a specific eye-tracking measure (including the one we discussed above) should be taken with a grain of salt.

be mutually exclusive, but an extended discussion about these competing accounts is beyond the scope of the present paper (see Ness & Meltzer-Asscher, 2017, for a more detailed discussion and an attempt to test these competing accounts).

Reduced effect of plausibility in longer dependencies

In addition to finding a plausibility effect in the critical region across all conditions, we also found that the plausibility effect which spilled over into the post-critical region was weaker in the Long conditions than in the Short condition. This suggests that the plausibility of a dependency had a weaker or less sustained effect when the dependency spanned a longer distance. We believe that this observation can be reconciled with the apparent lack of a plausibility effect in the active gap-filling regions in the Long conditions in W&P14's study. As we mentioned above, the fact that the effect of plausibility was weaker and less sustained in longer dependencies may have made it more difficult to detect in the Long conditions and could potentially explain why it was not detected in W&P14's study. In addition to the fact that W&P14's study had less statistical power than the present study, this difference could also be due to the use of self-paced reading in W&P14's study vs. eye-tracking in the present study. While the use of eye-tracking has allowed us to identify (i) a plausibility effect in the critical region across all conditions and that (ii) this effect was weakened in the post-critical region in the Long conditions, these effects may have been less localised and/or more varied in their timing in self-paced reading (Witzel et al., 2012), both of which could make the plausibility effect in the Long conditions more difficult to detect in W&P14's study. Further, the reduced effect of plausibility over longer dependencies is also in line with the results in another experiment reported by W&P14, which showed that the size of the filled-gap effect was numerically smaller in the two Long conditions compared to the Short condition.

We believe this may be a consequence of the demand of maintaining the filler in memory over a longer distance (Gibson, 1998) and/or the decreasing availability of a memory representation over time (McElree, 2000). Either of these factors may influence some aspects of language processing, even if they do not impede the active gap-filling mechanism. For example, we may take part of the effect of plausibility to reflect comprehenders' attempt to repair or reanalyse an implausible dependency. In turn, our observation that this effect is less sustained in the Long conditions may be taken to suggest that comprehenders are less likely (or make less effort) to repair an implausible dependency when they are faced with the demand of maintaining a filler in memory over a long distance and/or when the memory representation of the filler has undergone decay and become less available for retrieval over time. These findings add to the growing literature on how different aspects of language processing may be modulated by the distance between two elements in a dependency (e.g., Duffy & Rayner, 1990; Gibson, 1998; Lewis & Vasishth, 2005; Phillips, Kazaninaa & Abada, 2005). Future research will be required to better understand why dependency length may have differential effects on different linguistic processes.

Conclusion

The present study found clear eye-tracking evidence for active gap-filling in short and long dependencies alike. Our findings are in line with the long-standing generalisation that comprehenders build and interpret filler-gap dependencies predictively as long as they are grammatically licensed, but they diverge from the observations previously reported in self-paced reading study by Wagers and Phillips (2014). We propose that active gap-filling is fully at work regardless of dependency length, and that comprehenders can maintain the semantic features of a filler across long distance for active gap-filling. By increasing statistical power and recording comprehenders' eye-movements as they read naturally, the present study revealed a more detailed picture of how the processing of long-distance dependencies may be affected by dependency length.

Acknowledgements

We would like to thank Matt Wagers and Colin Phillips for sharing their stimuli and Rosanna Todd for her help with creating the new materials. We also thank Andrea Santi, Adrian Staub and two anonymous reviewers for comments on an earlier version of the manuscript.

References

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. http://doi.org/10.1016/j.jml.2007.12.005

Baddeley, A. (2010). Working memory. *Current Biology*, 20(4), R136-R140.

Barr, D. J., Levy, R. P., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. http://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. Journal of Statistical Software, 67, 1–48.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911. http://doi.org/10.3758/s13428-013-0403-5

Crain, S., & Fodor, J. D. (1985). How can grammars help parsers? In D. Dowty, L. Kartunnen, & A. Zwicky (Eds.), *Natural language parsing* (pp. 94-128). Cambridge: Cambridge University Press.

Duffy, S. A., & Rayner, K. (1990). Eye movements and anaphor resolution: effects of antecedent typicality and distance. *Language and Speech*, *33 (Pt 2)*(2), 103–19. http://doi.org/10.1177/002383099003300201

Fraser, B. (1971). A Note on the Spray Paint Cases. *Linguistic Inquiry*, *2*(4), 604–607. Retrieved from http://www.jstor.org/stable/4177680

Frazier, L. (1987). Sentence Processing: A Tutorial Review. *Attention and Performance XII The Psychology of Reading*. Retrieved from http://cnbc.cmu.edu/~plaut/IntroPDP/papers/Frazier87.sentProcRev.pdf

Frazier, L., & Flores d'Arcais, G. B. (1989). Filler driven parsing: A study of gap filling in dutch. *Journal of Memory and Language*, *28*(3), 331–344. http://doi.org/10.1016/0749-596X(89)90037-5

Garnsey, S. M., Tanenhaus, M. K., & Chapman, R. M. (1989). Evoked potentials and the study of sentence comprehension. *Journal of Psycholinguistic Research, 18*(1), 51–60. http://doi.org/10.1007/BF01069046

Gerlanc, D., & Kirby, K. N. (2015). bootES (Version 1.2). Retrieved from http://cran.r-project.org/web/packages/bootES/index.html

Gibson, E. (1998). Linguistic complexity: locality of syntactic dependencies. *Cognition*, *68*(1), 1–76. http://doi.org/10.1016/S0010-0277(98)00034-1

Jegerski, J. (2014). Self-paced reading. In J.Jegerski & B. VanPatten (Eds.), *Research mehods in second language psycholinguistics* (pp. 20-49). New York: Routledge.

Kirby, K. N., & Gerlanc, D. (2013). BootES: An R package for bootstrap confidence intervals on effect sizes. Behavior Research Methods, 45(4), 905–27. http://doi.org/10.3758/s13428-013-0330-5

Lee, M. W. (2004). Another look at the role of empty categories in sentence processing (and grammar). *Journal of Psycholinguistic Research*, *33*(1), 51–73. http://doi.org/10.1023/B:JOPR.0000010514.50468.30

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419. http://doi.org/10.1207/s15516709cog0000_25

Love, T., & Swinney, D. (1996). Coreference processing and levels of analysis in object-relative constructions; demonstration of antecedent reactivation with the cross-modal priming paradigm. *Journal of Psycholinguistic Research, 25(1),* 5–24.

McElree, B. (2000). Sentence comprehension is mediated by content-addressable memory structures. *Journal of Psycholinguistic Research, 29,* 111–123. https://doi.org/10.1023/A:1005184709695

McElree, B., & Griffith, T. (1998). Structural and lexical constraints on filling gaps during sentence comprehension: A time-course analysis. Journal of Experimental Psychology: Learning, Memory, and Cognition, 24(2), 432–460. http://doi.org/10.1037//0278-7393.24.2.432

Ness, T. & Meltzer-Asscher, A. (2017). Working memory in the processing of long-distance dependencies: Interference and filler maintenance. *Journal of Psycholinguistic Research, 46*, 1353-1365.

Omaki, A., Lau, E., White, I. D., Dakan, M. L., Apple, A., & Phillips, C. (2015). Hyper-active gap filling. *Frontiers in Psychology, 6*, 1–30. http://doi.org/10.3389/fpsyg.2015.00384

Phillips, C. (2006). The Real-time Status of Island Phenomena. *Language, 82*(4), 795–823.

Phillips, C., Kazanina, N., & Abada, S. H. (2005). ERP effects of the processing of syntactic long-distance dependencies. *Cognitive Brain Research*, *22*(3), 407–428. http://doi.org/10.1016/j.cogbrainres.2004.09.012

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*(6), 1401–1410.

Rappaport, M., & Levin, B. (1986). What to do with Theta Roles. (Lexicon Project Working Papers, 11). Cambridge, MA: MIT Center for Cognitive Science.

Santi, A., Friederici, A. D., Makuuchi, M., & Grodzinsky, Y. (2015). An fMRI study dissociating distance measures computed by Broca’s area in movement processing: clause boundary vs. identity. *Frontiers in Psychology*, *6*(May), 1–12. http://doi.org/10.3389/fpsyg.2015.00654

Stowe, L. A. (1986). Parsing WH-constructions: Evidence for on-line gap location. *Language and Cognitive Processes*, *1*(3), 227–245. http://doi.org/10.1080/01690968608407062

Sussman, R. S., & Sedivy, J. C. (2003). The time-course of processing syntactic dependencies: Evidence from eye movements. *Language and Cognitive Processes, 18*(2), 143–163. http://doi.org/10.1080/01690960143000498

Traxler, M. J., & Pickering, M. J. (1996). Plausibility and the processing of unbounded dependencies: An eye-tracking study. *Journal of Memory and Language*, *35*(3), 454–475. http://doi.org/10.1006/jmla.1996.0025

Tulving E. (1972). Episodic and semantic memory. In E. Tulving, W. Donaldson W (Eds.) *Organization of memory* (pp. 381–403). New York: Academic Press.

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. http://doi.org/10.1080/17470218.2013.850521

Wagers, M. W., & Phillips, C. (2009). Multiple dependencies and the role of the grammar in real-time comprehension. *Journal of Linguistics*, *45*, 395. http://doi.org/10.1017/S0022226709005726

Wagers, M. W., & Phillips, C. (2014). Going the distance: Memory and control processes in active dependency construction. *The Quarterly Journal of Experimental Psychology*, *67*(7), 1274–1304. http://doi.org/10.1080/17470218.2013.858363

Witzel, N., Witzel J., & Forster K. (2012). Comparison of online readinbg paradigms: Eye tracking, moving-window, and maze. *Journal of Psycholinguistic Research, 41*(2), 105-128. http://doi.org/ http://10.1007/s10936-011-9179-x