Predictions of listeners in response to speakers' repairs: Evidence from eye movements

3345233

MSc Psycholinguistics

The University of Edinburgh

2009

Abstract

There is now considerable evidence that upon hearing an utterance, listeners are able to make predictions about what is to follow. However, given the frequency of disfluency in normal speech we may wonder how this effects the predictions that listeners may make. While there is a growing body of literature concerned with how disfluencies may influence comprehension, there has been relatively little attention given to the case of repairs. The present paper presents an exploratory study using the visual world paradigm to investigate the predictions listeners make when speakers appear to change their mind while giving an instruction. We manipulated the lengths of the pauses and whether or not a retrace was present in the repair and found that this had an effect on both fixation likelihoods and their onsets. It is suggested that these findings may provide questions for future research, of which the visual world paradigm may continue to be a valuable tool.

1 Introduction

There is convincing evidence to suggest that prediction is involved in the process of language comprehension. Participants perform faster in lexical decision tasks when words are predictable given the context in which they appear (Schwanenflugel & Shoben, 1985). When reading, highly predictable words are also less likely to be fixated on than less predictable words (Ehrlich & Rayner, 1981).

Spontaneous human speech is rarely perfectly fluent, with one suggesting that six in every one hundred words spoken is effected by disfluency (Fox Tree, 1995), these may include fillers (such as *uh* and *um*), prolongations (for example pronouncing *the* as *thee* rather than the more common *thuh*), repetitions and repairs. Given the frequency with which disfluencies appear in natural speech and their ungrammatical nature, we may expect that they would present difficulty to the process of language comprehension. Yet, our ability to understand each other during everyday conversation provides a clear demonstration that the difficulty they pose is not insurmountable. How do listeners cope in the face of disfluent speech, and how do disfluencies interact with other processes involved in comprehension such as prediction?

For a listener trying to predict what is to come in an utterance, repairs pose a significant challenge. What predictions do they make about what is to come when the speaker appears to change their mind about what they are trying to say? In the present paper we will use one particular eye-tracking approach, the visual world paradigm, to investigate the effects of repairs on the predictions listeners make while hearing an utterance.

1.1 The visual world paradigm

Before summarising some of the previous work on both the production and comprehension of disfluencies, we will first examine the paradigm employed in the present study and in other work we will be discussed later. The visual world paradigm has provided a valuable on-line insight into language comprehension. In the earliest example, which came to define the paradigm, Cooper (1974) presented participants with a grid of images while they heard a story about a safari trip in Africa. When a lion was mentioned in the story, participants showed a tendency to fixate upon the image of a lion, rather than on images of other animals or objects. It is suggested here that the listener subjects the items in his or her visual field to the ongoing interpretations of an incremental comprehension system.

The paradigm was popularised by Tanenhaus, Spivey-Knowlton, Eberhard, and Sedivy (1995)

who used it to provide further evidence of the incremental nature of language comprehension (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995). Participants were shown a set of various shapes of different colours, some of which were marked by a star that was placed on them. While viewing these they were heard the instruction to *touch the starred yellow square*. Three different sets of shapes were used which corresponded to different conditions: whether the target was disambiguated by an early, middle or late word in the instruction. It was observed that the earlier in the instruction the disambiguating word appeared, the faster participants fixated upon the target, suggesting they were processing the sentence incrementally, rather than waiting for the instruction to be completed before reacting to it.

While the previous study shows participants identifying, via eye movements, the target of a sentence before its description was completed, the fact that the saccade appeared to be planned following the disambiguating word means that it is demonstrating prediction. A better example comes from Altmann and Kamide's (1999) study on interpretation of verbs. Participants saw an image of a scene containing a boy, a cake and various inedible objects, while hearing either *the boy will eat the cake* or *the boy will move the cake*. It was found that the onset of saccades towards the target were faster when the sentence contained *eat*, rather than *move*. This suggests that upon encountering the verb *eat*, listeners become aware of the selectional restrictions of the word, specifically that the theme must be edible, and predict that the theme will be the cake as it is the only edible object in the scene.

1.2 Production of disfluencies

Pickering and Garrod (2007) have suggested that a possible mechanism behind listeners' predictions during comprehension is through simulating their own production. If this is in fact true then perhaps people simulate their own disfluency when trying to process those of others (Watanabe, Hirose, Den, & Minematsu, 2008). Whether or not this is the case, the assumption that listeners consider past experiences of disfluency when processing disfluent material has permeated much research focussing on the comprehension of disfluencies. In light of this it is valuable to consider some of the findings from the production literature which highlights certain regularities in their occurrence, concluding with evidence that speakers may in fact use disfluencies to signal their difficulties to listeners.

Disfluencies appear to reflect problems with planning an utterance and the process of lexical access required to produce it. They tend to occur when speakers are unfamiliar with a topic

(Merlo & Mansur, 2004) but may also occur when the speaker is familiar with a topic, if the topic has large vocabulary rich with synonyms, such as those lecturing in the humanities (Schachter, Christenfeld, Ravina, & Bilous, 1991; Schachter, Rauscher, Christenfeld, & Tyson Crone, 1994). This finding reflects a general pattern that greater choice leads to a greater probability of disfluency (Christenfeld, 1994). Disfluencies also show a tendency to precede longer phrases (Shriberg, 1996) and new clauses (Clark & Wasow, 1998), as well as words with low contextual probability (Beattie & Butterworth, 1979) and images with ambiguous names (Schnadt & Corley, 2006).

Much thinking about one particular form of disfluency, repairs, has come to be influenced by Levelt's (1983) characterisation of the phenomena. Levelt identified three parts to a phrase which contains a repair. While some of these names do not appear frequently in the literature, and in other cases his terminology may altered in meaning, in the present paper we will follow his terms as they allow us to distinguish the many relevant parts of the repair.

The phrase begins with the *original utterance*, which continues until the *moment of interruption*. Within the original utterance lies the *reparandum*, the erroneous material, and may also include the *delay*, intended speech that lies between the reparandum and interruption. Following the interruption is the *editing phase*, which may contain an *editing term*, such as a filler and may also contain a silent pause. Finally, the *repair* is produced. In addition to the *alteration* itself, the repair may begin with a *retrace*, where the speaker repeats a portion of the original utterance which preceded the reparandum.

Earlier we reviewed some of the situations in which disfluency tends to appear, but why do speakers produce disfluencies rather merely remaining silent. One possibility is that disfluent material is in some way a by-product of the problems encountered by the production system. Perhaps fillers and prolongations are merely neutral noises generated by the production system when its normal operation has been interrupted as has been suggested by Levelt (1983), while repetitions may just be the production system repeating the most recent segment of an ongoing utterance while it tries to plan the next. In this case of repetitions we may ask why the material was repeated, rather than pausing. One answer may be that by pausing, the speaker leaves them self open to being interrupted by their listeners. A similar suggestion has been made elsewhere that filled pauses may perhaps be used by speakers to maintain their turn when in conversation (Maclay & Osgood, 1959).

The argument that disfluencies may be used as some form of signal has been particularly bolstered by two papers by Clark and Fox Tree (2002; Fox Tree and Clark,1997). Fox Tree and Clark (1997) examined a corpus for occurrences of the world *the*. For each instance the length

of the silence which followed was recorded. It was found that 81% of instances of the prolonged form of *the* were followed by a suspension of silence, compared to only 7% of instances of the word in its normal pronouncation. Fox Tree and Clark took this to suggest that the prolongation was being used by speakers to signal that a problem was upcoming.

(Clark & Fox Tree, 2002) took this idea further in their work with the fillers *uh* and *um*. The same corpus was examined and all instances of these fillers recorded. Again, attention was given to the lengths of the silence which followed their appearances. This analysis revealed that there were in fact differences between the lengths of the silences which occurred after each form of the filler. It was observed that the pauses that followed *um* tended to be longer than those which followed *uh*. With this finding, Clark and Fox Tree suggested that more than simply being signals of difficulty, the different forms of fillers actually had slightly different meanings, each suggesting whether the length of the upcoming delay was likely to be short or long.

For Clark and Fox Tree fillers are not merely a noise, a symptom of a problem an ongoing problem, rather they are a form of interjection, words in their own right, chosen by the speaker as they may choose any other word when planning an utterance. Unlike other words, fillers do not add to the meaning of the sentence, rather they are carried on what Clark (1996) calls the *collateral track*. This is where speakers are able to provide an ongoing commentary on their performance, helping to achieve successful communication.

The use of the London-Lund corpus in this study introduces a flaw which weakens their argument. In this corpus pauses were annotated using a system of dots and dashes corresponding to “one light foot” and “one stress unit” respectively. Clark and Fox Tree assigned a length of .5 units to each dot and then recorded pause lengths by counting the numbers of dots and dashes. The result is an arbitrary approach to measuring pause lengths, where actually accurately timing the lengths of the pauses may have been preferable.

We may also question Clark and Fox Tree’s decision to eliminate all speakers from their analysis who failed to produce more than one form of filler in the corpus. While we may understand this in context of the within-subjects approach they took to their statistics, it does raise several causes for concern. If individual fillers do indeed have different meanings then what does this say about speakers who choose to use only one. Are their pauses consistently of the same length? Is their “lexicon of disfluency” somehow lacking, leaving them unaware that they can make fine grain distinctions? Unfortunately, Clark and Fox Tree do not state the number of people eliminated for this reason, but we may well question whether the findings would be the same if they were included in the analysis.

We have presented several methodological reasons to be cautious about accepting Clark and Fox Tree's findings, competing evidence may not fare much better in the face of scrutiny. O'Connell and Kowal (2005) claim that not only do *uh* and *um* not differentiate in the lengths of upcoming delays but may fail to signal upcoming delays at all, and, unlike Clark and Fox Tree, they show this by actually recording the exact lengths of the pauses. The authors claim one of the strengths of their analysis being the choice of corpus, having used a selection of media interviews with Hillary Clinton. They argue that if disfluencies are actually used as signals, then highly experienced public speakers should be well skilled in their use. We would argue the opposite is equally plausible, highly capable public speakers may actually be less prone to being disfluent. In light of this, we would argue that their corpus has little value when trying to generalise to the population.

1.3 Comprehension of disfluencies

We will now review some of the literature which has investigated the response of listeners to the disfluent speech they encounter. In doing so we hope to not only provide a summary of some of the important findings, and a background to some of the methodological decisions made in the present study, but also to demonstrate that disfluent speech may provide a rich source of information that listeners appear capable of tapping into when engaging in communication.

In trying to determine how the listener copes in the face of disfluency, one possible explanation is that the disfluent material is in fact filtered out. This may leave only the fluent material which the comprehension system is already able to process. There is some evidence which suggests filtering may occur. Lickley (1995) played participants a disfluent recording instructing them how to build a model house and provided them with a transcript of the recording with the disfluencies cleaned up. Participants were then instructed to mark any point at which the recording and transcription differed, while actually following the instructions to build the house. It was observed that participants generally performed poorly at noticing discrepancies between recording and transcript. Bard and Lickley (1997) show this apparent "blindness" to disfluency may be longer lasting. Participants heard sentences containing repairs and were instructed to transcribe them verbatim. When their transcriptions were examined it was found that participants often failed to recall the reparandum in both repairs and, in particular, repetitions.

In contrast to these findings, there is also evidence which suggests that listeners may be aware of the disfluencies they encounter. Christenfeld (1995) played participants recordings of fluent and disfluent speech and subsequently asked them to estimate the number of filled pauses heard.

Unbeknownst to the listeners, some of the disfluent recordings had been edited, with the filled pauses removed and silent pauses of identical lengths left in their place. He found that the estimates increased when participants heard disfluent recordings, particularly those with the filled pauses remaining.

While it is difficult to directly compare Christenfeld's findings with Lickley's (1995), as Christenfeld investigated overall estimates of fillers rather than the specific instances of Lickley's, they may offer a possible explanation for the lack of recognition of disfluency. One variable manipulated by Christenfeld were the instructions participants received. They either received no instructions, or were guided to focus on the style of the speaker or the content of the recording.

Those participants who had been instructed to focus on the content performed poorly at estimating the number of filled pauses, while those who received no instructions did better, and perhaps unsurprisingly those told to focus on the style performing the best. As earlier mentioned, in Lickley's (1995) experiment, participants followed the recordings they heard to build a model house. By asking participants to treat the disfluent recordings as instructions they needed to follow closely, the experimenters were in effect pushing their participants to focus on the content. When we return consider this in terms of Christenfeld's (1995) findings, it is perhaps no surprise that participants appeared unable to notice discrepancies.

A possible explanation for Bard and Lickley's (1997) finding of poor recall for reparandum comes from the suggestion that the goal of comprehension is to arrive at a semantic representation of the utterance, rather than merely its surface form (cf. Bailey & Ferreira, 2003). As the reparandum may not be part of the intended semantic representation of the utterance, then it should be discarded as soon as the listener becomes aware, through the repair or repetition, of the true meaning of the utterance.

However, there is evidence that the reparandum may not be entirely discarded. Lau and Ferreira (2005) presented participants with disfluent sentences which included a main verb/reduced relative ambiguity. These sentences featured repairs immediately following a verb which was either syntactically ambiguous or unambiguously incompatible, with respect to the contents of the repair. When asked to make grammaticality judgements, participants rated the incompatible sentences as lower, suggesting an influence of the reparandum on the final interpretation of the sentence. While these findings reveal that the reparandum is not as cleanly discarded as thought, they do demonstrate that the disfluent material is retained in some form, suggesting it is not in fact being filtered out.

While our reviews of Clark and Fox Tree (2002) and O'Connell and Kowal (2005) suggests

that there is currently a lack of conclusive evidence about whether or not fillers are a signal, and if they are what exactly they may mean, there has been much interest in the effects hearing these "signals" may have upon listeners. Fox Tree (2001) observed that when performing an identical word recognition task participants recognition latencies were lower when the probe word was preceded by an *uh* than by silence created by the removal of a filler. Interestingly though the presence of an *um* prior to the probe word provided no benefit. This pattern was replicated using Dutch materials and listeners. These findings are perhaps confounded by silence length. In the conditions where fillers were removed, they were not replaced by silence. As a result the interruption prior to the probe word in the two edited conditions was on average 350ms shorter.

Fox Tree suggests that the *uh* provides a benefit by heightening the listeners attention to what is being said. Evidence for this claim has been provided by work with Event Related Potentials (ERP; Collard, Corley, MacGregor, & Donaldson, 2008). Participants heard recordings of speech, in some of which the final word had been digitally compressed, producing a poor telephone-like quality. It was expected that hearing this sudden change in quality would lead to a Mismatch Negativity (MMN), associated with detecting acoustic changes, and a P300, associated with re-orientating one's attention towards a novel item. In some of the sentences, an *uh* appeared immediately before the final word. As expected, in the fluent sentences the acoustically deviant final word led to an MMN and a P300, but where a disfluency was present the amplitude of both of these components were reduced. This suggests that the disfluency itself had already heightened attention towards the sentence, lessening the effect of the compressed final word.

Returning to Fox Tree's (2001) findings, we may ask why the *um* failed to heighten listeners attention to what is being said. Fox Tree's response is to suggest that her findings are consistent with those of Fox Tree (2001), providing a possible explanation that focussing attention may not be as helpful "when the length of the delay is indeterminant" (Fox Tree, 2001, p 325). While this appears to be a sensible suggestion, we may well question how much more determinant the length of the delay following an *uh* may be, as they are defined by Fox Tree (2001) as short and long, relative to only each other. Regardless of the explanation for these findings, the consequence has been a move towards only using *uh* in comprehension studies of fillers.

We began our review of the production literature by suggesting that listeners may react to disfluent speech by considering their own experiences of disfluency, either from hearing their own or others' disfluency. This hypothesis would suggest that on encountering disfluency listeners may anticipate its cause. As was previously mentioned, disfluency tend to precede words with low contextual probability. Listeners' expectations of this was investigated by Corley, MacGregor,

and Donaldson (2007). The N400 is an ERP component associated with experiencing difficulties integrating a word into the context a listener forms when hearing a sentence. The appearance of an unpredictable word during a sentence is likely to produce such difficulty. Corley et al. (2007) presented participants with sentences which ended with a word which was either predictable (e.g. "Everyone's got bad habits and mine is biting my nails") or unpredictable (e.g. "Everyone's got bad habits and mine is biting my tongue"). It was found that when participants encountered a filler immediately prior to the final word the amplitude of the N400 observed was less than when the sentence was fluent.

The previous finding suggests that listeners use the causes of disfluency in order to make predictions about nature of the material that follow the disfluencies they encounter. Further research using the visual world paradigm has also demonstrated that listeners are also capable of making predictions about the actual content which follows disfluencies. Research on production reveal that speakers are more likely to be disfluent when they are introducing new items to the discourse (Arnold, Wasow, Ginstrom, & Losongco, 2000), therefore upon encountering disfluency we may predict that the speaker is about to describe something new. Arnold, Fagnano, and Tanenhaus (2003) demonstrated that listeners can use disfluency to make predictions. Participants were shown a grid containing four objects and heard instructions to move the objects to different spaces in the grid. The set of objects in the grid contained the target, its cohort competitor (a word which shared the same initial phoneme) and two distractors. In each trial, participants heard a pair of instructions, examples of which are shown in (1). The first instruction contained either (1a) the target (candle) or (1b) the competitor (camel). While the second sentence always referred to the target, but was either (1c) fluent or (1d).

(1a) Put the grapes below the candle

(1b) Put the grapes below the camel

(1c) Now put the candle below the salt shaker

(1d) Now put theee, uh, candle below the salt shaker

Eye movement data revealed that upon encountering the disfluency in the second instruction, participants showed a tendency to fixate on whichever competitor had not been mentioned in the first instruction. This suggests that participants were aware that disfluencies tend to precede names which are new to the discourse and so predicted that they would be instructed to move the object which had not already been mentioned. A later study, also using the visual world paradigm, demonstrated that upon encountering disfluencies listeners would be more likely to

fixate on objects which were difficult to describe (Arnold, Hudson Kam, & Tanenhaus, 2007). While, when the second instruction was fluent, participants tended to gaze towards the previously mentioned object.

Studies, such as those previously reviewed, suggest that disfluencies such as fillers may fill what may appear to be a counter-intuitive role of helping listeners. They may help focus attention, helping to ease difficulty listeners may have faced in their absence. It is possible that there are some forms of disfluency which do pose a problem to listeners, and it has been proposed that repairs may be one of these forms (Fox Tree, 1995). Fox Tree had participants perform an identical word recognition task using materials which contained either repairs or repetitions prior to the probe word. Some of these sentences had been edited, with the disfluency replaced by a silence, so, for example "and the next figure, this has- it looks a little like a like a hammer" became "and the next figure, ... it looks a little like a like a hammer".

Fox Tree found that while repetitions appeared to be facilitating recognition, leading to faster reaction times when probe words were preceded by a repetition, repairs appeared to be detrimental, with reaction times slower when the repair was present than when it had been removed by silence. Repairs are not always harmful though, post hoc analysis revealed that false starts were only impairing recognition when they appeared in the middle of a sentence, rather than at the beginning.

In attempting to explain these findings Fox Tree presents a possible account of the experience of listeners on encountering a repair. When the listener becomes aware that an interruption has occurred they are forced to backtrack through what has already been said in order to determine the location of the error and connect what has preceded it to what will follow in order to form a coherent representation. With a repetition this is a simple process, listeners may just realise that the same thing has been said twice and discard one of the segments, but when it is a repair there is no clue of this sort to suggest where the reparandum began and what may be discarded. While people appear able to resolve the problems that arise from repaired speech, this takes effort which manifests as the slow down in reaction times.

While there have been many attempts to understand how listeners react to disfluencies such as fillers, there has generally been less attention given to repairs. The present paper seeks to redress this by investigating how listeners react upon encountering repairs, in particular we hope to observe the effects the lengths of the pauses that appear during the editing phase of a repair, and whether or not the speaker produces a retrace, have on the predictions listeners make.

This is an exploratory study and as such we will forgo making any predictions about what

may be observed, but the literature suggests several findings that may arise. In Fox Tree's (1995) account of repairs she suggests that listeners may have to backtrack upon encounter a repair, and we see no reason why the speaker them self does not perform the same process. If backtracking is cognitively demanding then we may expect that a speaker may pause longer the further back into an utterance they are forced to go. This may lead listeners to predict that with longer pauses, a listener is repairing an earlier part of the utterance than a short pause may suggest. If this pattern does hold then perhaps when a speaker is seen to behave incongruently, for example by producing a long pause and then only backtracking one word into the original utterance, a listener will react differently.

Another question which could be explored comes from other work on how well listeners cope with repairs. Howell and Young (1991) asked participants to rate sentences containing repairs according to how easy they were to comprehend. From this it emerged that listeners have a preference for repairs immediately preceded by a pause, rather than other sentences where the pause appears elsewhere. If this is indeed the case then we may wonder what effect the presence of a retrace has upon listeners reaction to repairs.

2 Experiment

Participants were presented with auditory instructions to click on specific shapes displayed on a computer screen. While these sentences were heard, their eye movements were recorded with the use of an eye-tracker. In some of these sentences a repair appeared to be made, the edit phase of which always contained the filler *uh* followed by a silent pause. Two details of these repairs were manipulated: firstly, the length of the pause was either short or long; secondly, whether or not an additional adjective was spoken before the disambiguating adjective, forming a retrace.

2.1 Participants

Twenty four native British English speaking students (6 male, 16 female; ranging between 18-34 years old) from the University of Edinburgh were paid to participate in the study. All participants had normal, or corrected-to-normal, vision.

2.2 Procedure

Participants were seated in front of a 21" CRT computer screen, at a resolution of 1024x768 and a refresh rate of 100Hz. Eye movement data was recorded by an SR Research EyeLink 2000, sampling at 1000Hz on the right eye (although viewing was binocular). Auditory stimuli were played via a pair of speakers located behind the participant. The experiment was designed and controlled using SR Research's Experiment Builder software. Instructions displayed on the screen prior to the experiment informed participants that they would see sets of shapes and would hear sentences, recorded in a previous experiment, instructing them to click on particular shapes, and they should follow these instructions. A practice trial was carried out before the experimental block began. Participants were shown an image of eight shapes and immediately following its appearance they heard a sentence instructing them to click on one of them.

Between the practice trial and the beginning of the actual experimental block, the eye-tracker was calibrated using the EyeLink calibration routine. Participants saw a series of nine small circles appear on the screen and were instructed to gaze at each, once the participant's fixation had been recorded the next circle would appear and they would gaze at that leading to the appearance of the next, and so on. This process was repeated in order to validate measurements, and if necessary the experimenter would repeat the calibration routine. When the eye-tracker was suitably calibrated the experiment began.

Each trial began with a fixation cross in the middle of the screen, allowing the experimenter to ensure the eye-tracker was correctly calibrated. Upon the participant gazing at the cross, the experimenter triggered the appearance of each image, presented alongside the appropriate sentence. The experiment consisted of sixty-four trials, broken down into four blocks of eight, providing participants with an opportunity to rest between blocks. At the end of each rest period, the measurements were validated and, if necessary (or if the participant had moved during the break), another calibration routine was run. The experiment lasted fifteen minutes.

2.3 Auditory and visual stimuli

Images of four shapes (square, circle, triangle and star) were created using vector graphics software. Two versions of each were created, one with stripes and the other with spots. Each of these shapes were copied and recoloured using four different colours (red, yellow, green and blue), producing thirty-two images in total. These images were assembled in various combinations to create sixteen sets of eight. All of the sets contained two images of each colour, these pairs of images

were always of the same shape, with one always having spots, while the other had stripes.

Sentences were created to instruct participants as to which shape they were to click on in each trial. In total, sixty-four sentences were created, of which twenty were experimental items, containing repairs, while the remaining forty-four were filler sentences. All sentences began with *Click on the...*, followed by the the shape participants were to click on. Each shape was identified using two adjectives, the colour and the pattern, either stripy or spotty, with the adjectives always appearing in this order. A female native speaker of British English was recorded repeating these sentences in a studio at the University of Edinburgh, at a sampling rate of 46kHz. Sentences used in experimental trials included a repair, following the second adjective, which was scripted for the speaker, they were instructed to include an *uh* in the edit-interval, and follow this with a pause. The speaker was advised that the length of the pause was not important and instead they should ensure that the filler sounded natural and that they spoke at a slow, but plausible, rate. Each sentence was recorded three times, with the most natural sounding version used in the experiment.

Experimental items were manipulated to create four types of five sentences as shown in (2). The four conditions were (2a) a short pause following the filler (denoted by [s]) with a one adjective-long repair, (2b) a short pause followed by a two adjective-long repair, (2c a long pause (denoted by [l]) followed by a one adjective-long repair, and (2d) a long pause followed by a two adjective-long repair. In all cases only the second adjective, relating to the pattern of the shape, was erroneous and needed correction.

(2a) Click on the red stripy uh [s] spotty square

(2b) Click on the red stripy uh [s] red spotty square

(2c) Click on the red stripy uh [l] spotty square

(2d) Click on the red stripy uh [l] red spotty square

In order to reduce the risk of participants assuming that fillers would always be followed by a repair, several types of disfluent filler sentences were created, in addition to twenty-two fluent instructions (3a). To suggest that a filler did not always mean a mistake had been made, six filler sentences included repetitions of either one adjective (3b) or two (3c), preceeded by a filler and pause. While these filler sentences suggest that fillers do not always precede repairs, they do share, with the experimental items, the presence of additional adjectives. To prevent this assumption from being made, two additional sets of six filler sentences were included, suggestive of disfluency symptomatic of lexical selection problems (Schnadt & Corley, 2006). In the first set (3d), a filler and pause was inserted between the two adjectives of an otherwise fluent sentence,

while in the second set (3e), the filler and pause were inserted between the second adjective and the name of the shape.

In each set of disfluent filler sentences, half of the sentences contained a short pause, while the remaining half used a long pause. For both the experimental and filler sentences, a constant length of silence was used for each pause. A time of 833ms was used as a long pause. This figure came from Fox Tree (1995), who found that this was the mean length of the pauses in repairs from a corpus of spontaneous speech. To provide short pauses, this figure was roughly halved, giving a pause of 416ms. The silences were taken extracted moments of silence in the recording process and inserted in the sentences in place of any natural pause. The mean length of the filler in the experimental sentences was 501ms ($SD = 68.19$). A paired samples t-test showed there to be no significant difference between the mean lengths of the fillers in either the short pause or long pause conditions ($t(9) = .345, p > .1$) suggesting no risk that any findings obtained may be confounded by the lengths of fillers. This is particularly important given Bailey and Ferreira's (2003) suggestion that it is the interruption that fillers provide that drive their effects on comprehension. A second paired samples t-test showed that the total length of the edit-interval (the filler and pause combined) differed significantly between conditions ($t(9) = -10.985, p < .001$).

(3a) Click on the red spotty square

(3b) Click on the red spotty uh [] spotty square

(3c) Click on the red spotty uh [] red spotty square

(3d) Click on the red uh [] spotty square

(3e) Click on the red spotty uh [] square

2.4 Analysis

The data recorded by the eye-tracker was analysed as follows. As our auditory stimuli varied in length, we could not simply compare eye movements across time. To solve this four epochs were created, each of varying lengths both relative to each other and between each sentence. Each epoch contained a particular segment of the sentences: the first containing the beginning of the sentence up until the end of the reparandum (*Click on the red stripy*), the second containing the filler (*uh*), the third consisting of only the pause, and finally the fourth which ran from the repair to the moment the participant brought each trial to an end by clicking on a shape. For each sentence the lengths of these segments were measured. During each trial a message was written to the

EyeLink output file at the moment when each sentence began. Using this message and the length of each segment obtained earlier, new messages were written into the output file to signal the beginning of each epoch. The mean lengths of each epoch are shown in Table 1

Table 1: Mean duration (in milliseconds) of epochs in auditory stimuli (standard deviations in brackets)

	Reparandum[1] Click on the red stripy	Filler uh	Pause[2] []	Repair[3] spotty/red spotty square
Short pause/One-word repair	1947.8	510.6	416.0	1943.6
	(128.3)	(74.5)	—	(252.0)
Short pause/Two-word repair	1916.5	497.0	416.0	2080.3
	(172.4)	(73.0)	—	(269.6)
Long pause/One-word repair	1950.8	496.0	833.0	1831.0
	(289.8)	(87.5)	—	(198.4)
Long pause/Two-word repair	2001.4	511.6	833.0	2373.6
	(336.2)	(60.7)	—	(864.3)

The size and position of each shape was determined and entered into the EyeLink output file as interest areas and matched with the appropriate images. These interest areas allowed us to obtain a record of all eye-movements on the various shapes within each image. Using EyeLink Data Viewer, data of all fixations on each shape was extracted for each individual epoch.

Each shape was assigned to one of three categories, corresponding to different regions of the image: the target; the decoy, the erroneous target of the instruction; and distractors, those shapes which were not mentioned in the instruction. For each epoch, the first fixation in each region was used in our analysis. Where a fixation made in a previous epoch was maintained in the current this fixation was ignored and the next fixation made within the epoch used instead. Where there was no second fixation, this was treated as a case where no fixation had been made.

For our analysis we focused individually on each of the three regions. Within each epoch analyses were run on both fixation likelihoods and onsets of fixations. As likelihood is a binomial variable, fixation likelihood was modelled with the use of logit mixed effects models. While an alternative approach may be perform an ANOVA on arc-sine transformed data, it has been suggested that this may lead to spurious results (cf. Jaeger, 2008). Mixed models also remove the need for separate by-participants and by-items analysis, by allowing participants and items to be entered into models as random effects.

[1]Measured from onset of sound file, includes any initial silence
[2]In each condition the length of the pause was constant
[3]Measured from onset of repair until the participant clicked upon a shape

In all epochs our analysis followed the same pattern, irrespective of the dependent variable used. Firstly, a null model was created which included an intercept and the participants and sentences as random effects. This model is equivalent to saying that any significant differences are due solely to individual differences of participants and a, as yet unaccounted for, difference in sentences.

Secondly, "full" models were created which contained fixed effects. This was done by incrementally adding relevant predictors relating to our conditions. The log-likelihoods of these full models and the null models were then compared using log-likelihood ratio tests, $-2(l_1 - l_0)$, with a χ^2 test assessing whether the inclusion of a variable was significantly improving our model, with those variables which failed to improve the model being removed. In constructing models to predict onsets of fixations linear mixed effects models were used, following the same process as was used to select the best fitting logistic mixed effects models with one exception: Markov Chain Monte Carlo sampling performed over $10,000$ simulations in order to estimate coefficient probabilities.

3 Results

Due to problems with the PC controlling the presentation of the experiment, one participant had to be removed from the analysis. Additionally, as a result of experimenter error all data associated with one experimental item (from the long pause/retrace condition) had to be removed. The accuracy of participants in clicking on the correct shape was not recorded, but any trial where a participant clicked on a shape prior to the instruction ending was treated as an error and ignored.

As the first two epochs (the original utterance, and the editing term) of the stimulus precede the appearance of the pause and retrace which generate the conditions of the experiment, using either the length of the pause or the presence of a retrace as a predictor of fixations, or their onsets, would be meaningless. In light of this, we will not report any statistics performed with data recorded in these epochs. Mean probabilities of fixation likelihoods in these epochs, and in the pause and repair epochs are shown in Figure 1. All analyses were performed in R (R Development Core Team, 2008), using the lme4 package (Bates, Maechler, & Dai, 2008).

3.1 Pause

While the duration of the pause, and as a result this epoch, was an experimental manipulation, it immediately preceded the beginning of any possible retrace and so it would be meaningless to

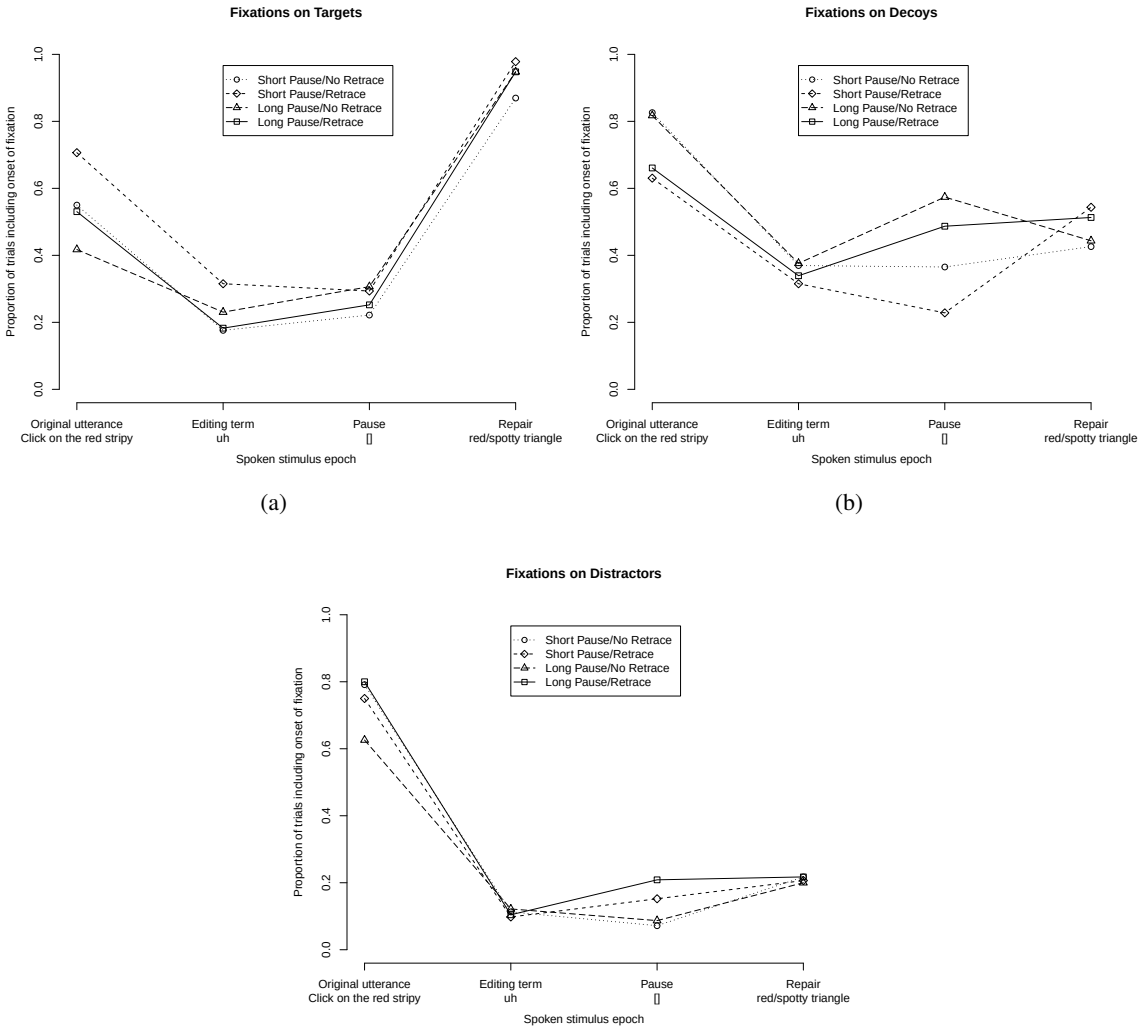

Figure 1: Probabilities of fixating upon either the (a) Target, (b) Decoy, or (c) Distractor items during each epoch of the spoken stimulus.

enter the presence of the retrace as a predictor in our models. Only the length of the pause (coded as either 0 for short, or 1 for long) were tested. Coefficients for all fixation likelihood models which improve fit are given in Table 2.

3.1.1 Fixations on targets

In the period immediately following the epoch, the null model of fixation likelihood was not improved by the addition of pause length ($\chi^2(1) = .19, p > .05$). However, the best fitting model for the onset of fixations was found to include the length of the pauses as a predictor ($\chi^2(1) = 20.42, p > .001$), with a log-likelihood of -583.9. This suggested that participants tended to fixate later when the pause was longer.

There is a obvious problem with this finding as a result of the length of the epoch being dictated by the condition. When the pause was longer, the epoch was longer and therefore it is not surprising that there would be more later fixations. In order to test this a new epoch was created for the experimental items containing a short pause. For these items the epoch was extended, eating up 417ms of the following. This allowed us to compare eye-movements in the 833ms following the end of the editing item, regardless of condition. With this new epoch it was observed that pause lengths were no longer predicting onsets of fixations, providing no improvement to the null model ($\chi^2(1) = 1.19, p > 0.05$).

3.1.2 Fixations on decoys

When fixation likelihood for decoy items was examined, it was observed that addition of pause length improved the null model ($\chi^2(1) = 15.66, p > .001$), with a log-likelihood of -280.1. Pause length was also found to be predicting the onsets of fixations ($\chi^2(1) = 8.23, p > .01$), with log-likelihood -1225. With longer pauses, participants appeared more likely to fixate, and when they did they were quicker to do so. Given what has already been said about the varying lengths of this epoch, we considered these findings using our newly extended epoch. As before, the effect of pause length on fixation onsets disappeared ($\chi^2(1) = 0.03, p > .05$), but the increase in fixation likelihood given short pauses remained ($\chi^2(1) = 8.83, p < .01$), with log-likelihood -574.3.

3.1.3 Fixations on distractors

While pause length provided no improvement to the null model of fixation likelihood ($\chi^2(1) = 1.23, p > .05$), it again appeared to be likelihood when fixations upon distractor shapes were

analysed ($\chi^2(1) = 14.73, p < .001$), with log-likelihood -362.6. As in the previous two cases though, this improvement was not present in the modified epoch ($\chi^2(1) = 0.12, p > .05$)

3.2 Repair

We will now turn to the final epoch, the repair. In addition to pause length, we may now enter the presence of a retrace into our models (coded as 1, for the presence of one adjective, and 2, for the second adjective which forms the retrace). If either predictor fails to improve upon the null model then it may be removed, however if both predictors improve the model then, as the models share the number of degrees of freedom, we will use whichever model provides the log-likelihood closest to zero. The presence of a retrace in two of the conditions, adding material between the beginning of the epoch and the disambiguating adjective that all items share, creates a problem of how to deal with the onset times for fixations. Before offering our posthoc solution for this problem, we will focus on whether or not fixations are likely to occur.

3.2.1 Fixation likelihoods

From the beginning of the repair, the best fitting model of fixation likelihood upon the target shapes contained an interaction between pause length and the presence of a retrace, ($\chi^2(1) = 4.65, p < .5$) with log-likelihood -97.04. When the pause length was long and the repair contained a retrace participants were more likely to fixate on the target. In the cases of both fixations upon the decoy ($\chi^2(1) < .01, p > .05$) and upon the distractors ($\chi^2(1) = 0.01, p > .05$), neither adding pause length nor the presence of a retrace was found to improve upon the null models.

Table 2: Coefficients and probabilities for fixation likelihood models in each epoch

	Shape	Predictor	Coefficient Estimate	Std. Error	p
Pause	Target		No improvement on null model		
	Decoy	intercept	-0.4888	0.1759	$< .01$
		Pause Length	0.5048	0.1630	$< .01$
	Distractor		No improvement on null model		
Repair	Target	intercept	2.2477	0.3695	$< .001$
		Pause Length	1.0925	0.5379	$< .05$
		Retrace	2.0394	0.8194	$< .05$
		Pause Length $*$Retrace	-2.0394	1.0358	$< .05$
	Decoy		No improvement on null model		
	Distractor		No improvement on null model		

3.2.2 Onset of fixations

Depending on whether or not an item appears in a retrace condition, the onset of the disambiguating onset may not coincide with the beginning of the epoch. The result of this is that onset times are not locked to a particular moment in the time course of each sentence. In order to get past this problem the span of the fourth epoch was reduced so that it would commence with the onset of the disambiguating word, irrespective of what preceded. All fixations appearing in this narrower epoch had their onsets adjusted so they would be relative only to this shared epoch.

Following these adjustments the best fitting model of predicting onsets for both the target ($\chi^2(1) = 9.38, p < .01$) and the decoy ($\chi^2(1) = 5.23, p < .05$) included only the presence of the retrace, with log-likelihoods of -2886 and -1304 respectively. Neither predictor was able to improve the null model for onset of fixations on distractors ($\chi^2(1) = 0.10, p > .05$). Coefficients for these models appear in Table 3.

Table 3: Coefficients and probabilities for best fitting models of fixation onsets in the Repair epoch

	Predictor	Coefficient Estimate	95% CI (lo)[4]	95% CI (hi)[4]	p[4]
Target	intercept	709.0	629.1	789.35	$< .001$
	Retrace	-124.7	-201.9	-42.87	$< .001$
Decoy	intercept	709.0	630.4	789.43	$< .001$
	Retrace	-124.7	-206.0	-45.91	$< .001$
Distractor	No improvement on null model				

4 Discussion

Analysis of the eye movement data recorded while participants heard instructions containing repairs has led to the emergence of several findings. In the pause epoch, participants showed a greater likelihood to fixate upon the decoy shape when the pause following the filler was short. The length of this epoch varied depending on the condition each sentence was in, raising the possibility that our finding was confounded by epoch duration. In post hoc tests, the length of this duration was adjusted to be uniform across conditions and the significant finding remained.

In the repair epoch an interaction was found between pause length and the presence of a retrace which appeared to be driving fixations on the target shape. Specifically, a short pause and the presence of a one word retrace lead to an increase in fixation likelihood. In order to allow us to make comparisons of fixation onsets between conditions at this point we had to remove the

[4]Estimated using $10,000$ Markov Chain Monte Carlo Simulation. p values for coefficients used in linear mixed effects models should be considered "anti-conservative"

retrace from the epoch to ensure that all epochs began with the onset of the alteration. When these were examined, it was found that the presence of the retrace was associated with later onsets on both the target and decoy, although not the distractors. While none of the findings observed directly contradict each other we do not believe that together they tell a coherent story of listener's predictions. As a result we will consider each finding independently of one another.

Earlier we considered what effect any backtracking a speaker may have to do when they are forced to make a repair may have on the pauses they produce. One suggestion was that there may be a direct link between the amount of backtracking a speaker does and the length of their pause, as a result of increased cognitive burden. We asked if any incongruency between pause length and the length of a repair would have an effect on the listener. No such effect was observed, however it does appear that when the lengths are congruent listeners are more likely to fixate on the target, reflected in the interaction found in the repair epoch.

Without a similar interaction for the onsets of fixations or any sign that incongruency is detrimental to the listener it is not possible to be entirely conclusive about what this finding means. Despite this, the finding does provide suggestions for future research. Using a similar paradigm to the present study with longer pause lengths and longer retraces this effect could be further investigated, in addition listeners predictions with repairs may be compared to predictions they may make with repetitions, where it is possible that more backtracking occurs.

While we may presume that an increase in fixation likelihood on the targets would be matched by faster fixation onsets, it was found that the presence of a retrace was actually having the opposite effect. As with decoys, where material was present prior to the alteration, fixations on the target produced later onsets. While this appears incompatible with the finding just described, it appears compatible with one of the findings of Howell and Young (1991).

They observed that when participants were shown sentences containing repairs and repetitions, they considered those with a pause immediately before the onset of the alteration as easier to comprehend. One might suggest that our participants' later onsets were a symptom of the difficulty that leads listeners to rate sentences such as these as harder to understand. However, Howell and Young had other participants repeat these sentences without the disfluency, after hearing them produced by a speech synthesizer (Experiment 2b). Using initiation of speech as a measure, they revealed that participants performed better when the pause was prior to a retrace.

As suggested earlier, we do not believe that our findings come together to form an account of the predictions listeners make when encountering speakers making repairs. While future work may find links between the various effects we observed, we view them currently as a set of isolated

findings. Our intentions for this the present study was to explore the issue in order to generate hypotheses for future research, and we believe our findings raise questions which others may begin to answer. In the remainder of this paper we will consider some future research which may be carried out, identify some weaknesses in our own methodology and analysis, and evaluate the paradigm used.

Given that long pauses increase fixation likelihood upon decoys in the pause epoch, and then, in the subsequent epoch, help to drive fixations towards the target we may wonder about the role the retrace plays in shifting the focus of the fixations. One possible question that may be asked is whether it is the content of the retrace or merely the extended pause it provides prior to the alteration. Further experiments could manipulate the form of the retrace further by replacing it with with new words that did not appear in the original utterance or with silences of matched lengths.

Similar questions have already been asked of fillers. It has been suggested by Bailey and Ferreira (2003) that some of the benefits that fillers appear to offer are not in fact due to anything in their nature, but merely the interruption they provide. Bailey and Ferreira investigated whether or not disfluencies could be processed by the parser by examining their effect on garden path ambiguities (Experiment 1). The role of fillers was tested with the head noun position effect. This effect says that given an ambiguous structure such as “While the boy scratched the dog”, readers will find it harder to process when a modifier appears after the ambiguous noun (e.g. “the dog that was hairy”), but a modifier prior to the noun (e.g. “the hairy dog”) will make no difference.

Two forms of garden path sentences where used to create experimental items: subordinate-main ambiguities (e.g. “While the man hunted the deer ran into the woods”) and coordination ambiguities (e.g. “Sandra bumped into the busboy and the waiter told her to be careful”). For each form of ambiguity, five types of sentence were created: an unmodified form, a prenominal modifier, a postnominal modifer, and two disfluent sentences with two fillers either prenominally or postnominally. When participants heard these sentences they rated those with a postnominal disfluency as less grammatical, matching the same pattern of judgements as were made with modifiers. These findings suggest that fillers may be parsed.

When replacement of disfluencies by environmental noises (Experiment 2), such as door bells ringing, produced similar findings, the possibility arose that there was nothing special about fillers that allow them to exhibit the head noun effect. Clark and Wasow (1998) show that fillers tend to appear at the beginning of new clauses, if listeners are aware of this then the presence of a filler may influence a listeners syntactic interpretation of a sentence. With a coordination ambiguity,

a filler prior to the second noun phrase (following *and* in the example above) may provide a "good signal" to listeners by highlighting the end of the previous constituent (preceding the *and*), whereas a filler in the first noun phrase (e.g. before *waiter*) may provide a "bad signal".

When participants heard these sentences (Experiment 3) and others with modifiers in similar locations, it was found that good signals led to higher ratings of grammaticality that bad signals, with this pattern reversed when fillers were replaced by modifiers. When modifiers were replaced by environmental noises (Experiment 4) the same pattern was found between those sentences with disfluencies and those with noises. Given the finding that noises have the same effect as fillers, it has been suggested that what is beneficial about fillers in cases such as these is rather to do with the interruption and additional opportunity to consider the utterance they provide than anything specific to the sound of a filler. However, Bailey and Ferreira's (2003) disfluent materials are generally considered to be of poor quality, so we may question how representative they are of the disfluencies listeners encounter in spontaneous speech.

We believe that our paradigm may be able to further investigate this idea. By comparing the predictions listeners make when they hear a repair containing an editing term, by those predictions where the term is replaced by a silence or noise of similar length we may be able to learn more about the exact nature of the benefits fillers provide. Furthermore, using *um*, in addition to *uh*, while maintaining our manipulation of short or long pause lengths may tell us more about this but also provide evidence of whether listeners are sensitive to the lengths of the delays which (Clark & Fox Tree, 2002) suggest different fillers signal. We propose that if listeners are aware of this distinction, then upon encountering incongruent pauses (an followed by a long pause, and vice versa) their predictive behaviour may show changes. Similar to the possible effects of incongruency on backtracking and pause length suggested earlier.

As with retraces, we may also ask what effect delays (in the Levelt sense, correct material following the reparandum) have upon the listener's predictions. When a delay occurs, listeners have no – currently understood way – of knowing that this is in fact a delay, rather than a continuation of the reparandum, until they become aware that it is being repeated following the alteration. By altering the images used in our sets so that there is more than one shape which appears in different colours, each instruction would only become disambiguated when the name of the shape had been used, while with our sets a shape is disambiguated when the pattern is named.

If the original utterance contained the full name of a shape, for example "red stripy square", and the speaker made a repair then the ability to distinguish between what is a reparandum and what is a delay may likely lead to faster onsets of fixations on the target. Returning to our example,

if the listener somehow became aware that *square* was a delay, then upon hearing the alteration of the pattern name listeners would be able to accurately predict the target. While we have no theory of if, and how, listeners may differentiate between reparandum and delay, one possibility may lie in Shriberg's (2002) observation of changes in the tone of speech during a reparandum, although she provides no evidence of whether this tone is continued during the delay.

We will briefly highlight several methodological flaws within our paradigm which may be corrected in any future research. While the speaker was naive to the aims of the study and was instructed to produce speech that sounded as natural as possible, no independent ratings were performed on the auditory stimuli used. As a result, we are unable to provide evidence that all materials sounded as natural as intended. However, following debriefings, no participant suggested that they felt the stimuli had been edited in any way. In addition, while we see no reason in the present study as to why this should be of particular concern, the failure to counterbalance materials was an omission that should be instated in any further work with this paradigm.

Finally, we would advice caution in interpreting our findings from the pause epoch. In order to ensure the duration of the pause epochs remained uniform, whilst maintaining a meaningful an actual distinction between conditions, some data points included in the short pause conditions for that epoch may have occurred following the onset of the repair phase, while data points included in the long pause condition had all begun prior to this repair. As a result we are unable to account for the influence of the repair upon half of the data appearing in this epoch. However, we do not believe a similar concern should be taken with the fixation onset findings in the repair epoch as redefining this epoch instead involved narrowing the duration, and as a result excluding, rather than including, additional data.

We will finish by consider the validity of experiments of this sort, beginning with a concern of our own, followed by a concern expressed elsewhere that believe evidence suggests is not in fact a problem. It is often cited that disfluencies occur at a rate of six words for every one hundred spoken. However for listeners who find themselves participating in disfluency experiments the rate they encounter may be vastly different. In the present study, over 50% of auditory stimuli heard by participants contained a disfluency.

While the desire to not bombard participants with fillers means that often in psychological experiments, participants repeatedly encounter particular phenomena at an unnatural rate we must ask how appropriate this is for experiments which are concerned with predictions that are made. A participant who encounters disfluent material almost ten times more often then they may expect to in the real world may rightly assume that these are significant and begin to give them extra

attention. It is one of the perils of experimental research that participants may discover patterns in our materials, however we must ask how suitable this is in cases where our experiments are designed to investigate the predictions listeners make based on patterns that occur in the real world.

In designing materials for disfluency experiments, we may "muddy the waters" with the use of fillers items. Fillers may be inserted in various locations differing from where they are placed in the experimental items, other forms of disfluency may be used. However as our summary of the production literature suggests, there are many different causes of disfluency and we cannot hope to create a filler for every possible situation. Further, we must ask what effect our filler items are having. When we are devising fillers which hide the patterns which we hope to study, then are we risking sending participants the message that it is not worth making predictions about disfluencies, expecting participants to make real world predictions in a context which is statistically different from the real world.

We do not believe that this should curtail research into disfluency, and we would in fact suggest that online measures such as the visual world paradigm may offer particular immunity to these problems, by tapping directly into processes which are perhaps not as heavily influenced by short term experience. We would, however, raise a note of caution in deciding the form and distribution in which disfluencies may take when assembling materials.

One criticism which has been pointed at the use of the visual world paradigm in comprehension research is that it presents an unnatural situation for the listener (e.g. Corley et al., 2007). Listeners outside of the laboratory are rarely presented with a finite set of potential referents for the spontaneous speech encounter. As is often the procedure of a visual world paradigm experiment, listeners are able to preview these potential referents, providing them with the opportunity to access the names of each of these and integrate these into a map of the scene (Dahan & Tanenhaus, 2005). While in the natural dialogue, listeners may found themselves surrounded by possible referents, the limited range and the previews the visual world paradigm offers perhaps provides an unnatural advantage.

There is, however, evidence of eye movements which are not mediated by the scene participants view (e.g. Altmann & Kamide, 1999; Dahan, Magnuson, & Tanenhaus, 2001; Dahan, Magnuson, Tanenhaus, & Hogan, 2001). Dahan, Magnuson, and Tanenhaus (2001) presented participants with scenes containing triplets of objects, one of which was the target (*bench*), and a phonologically unrelated distractor (*lobster*). These three words shared a similar phonological form, however while one competitor of the target was a high frequency word (*bed*), the other was

a low frequency word (*bell*). Participants then heard sentences such as "Pick up the bench".

It was found that before the offset of the instruction, when the phonological form the target would take was still ambiguous, participants showed a tendency to fixate on high frequency competitors rather than low frequency competitors. On previewing the scene, the names of all items were likely accessed, however this finding suggests that behaviour may still be mediated by factors removed from simply the presentation of these items.

The present study highlights the insight which the visual world paradigm may bring to disfluency research. The findings observed, particularly an effect of congruency between the lengths of pauses and the presence of a retrace driving fixations upon the target during the repair epoch, raise new questions which may be investigated further. In this respect, while our findings may fail to provide a coherent picture of listeners' predictions upon encounter a repair they do offer individual avenues for future research, using both the visual world paradigm and other approaches.

References

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, *73*(3), 247–264.

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal theee, um, new information. *Journal of Psycholinguistic Research*, *32*, 25–36.

Arnold, J. E., Hudson Kam, C. L., & Tanenhaus, M. K. (2007). If you say *thee uh* you are describing something hard: The on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 914–930.

Arnold, J. E., Wasow, T., Ginstrom, R., & Losongco, T. (2000). Heaviness vs newness: The effects of structural complexity and discourse status on constituent ordering. *Language*, *76*, 28–55.

Bailey, K. G. D., & Ferreira, F. (2003). Disfluencies affect the parsing of garden-path sentences. *Journal of Memory and Language*, *49*, 183–200.

Bard, E. G., & Lickley, R. J. (1997). On not remembering disfluencies. In *Proceedings of the international conference on spoken language processing.*

Bates, D., Maechler, M., & Dai, B. (2008). lme4: Linear mixed-effects models using S4 classes [Computer software manual]. Available from `http://lme4.r-forge.r-project.org/` (R package version 0.999375-23)

Beattie, G. W., & Butterworth, B. (1979). Contextual probability and word frequency as determinants of pauses in spontaneous speech. *Language and Speech*, *22*, 201–211.

Christenfeld, N. (1994). Options and ums. *Journal of Language and Social Psychology*, *13*, 192–192.

Christenfeld, N. (1995). Does it hurt to say um? *Journal of Nonverbal Behavior*, *19*, 171–186.

Clark, H. H. (1996). *Using language*. Cambridge, MA: Cambridge University Press.

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*, 73–111.

Clark, H. H., & Wasow, T. (1998). Repeating words in spontaneous speech. *Cognitive Psychology*, *37*(3), 201–242.

Collard, P., Corley, M., MacGregor, L. J., & Donaldson, D. I. (2008). Attention orienting effects of hesitations in speech: Evidence from ERPs. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 696–702.

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*(1), 84–107.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition*, *105*, 658–668.

Dahan, D., Magnuson, J., & Tanenhaus, M. (2001). Time course of frequency effects in spoken-word recognition: Evidence from eye movements. *Cognitive Psychology*, *42*(4), 317–367.

Dahan, D., Magnuson, J., Tanenhaus, M., & Hogan, E. (2001). Subcategorical mismatches and the time course of lexical access: Evidence for lexical competition. *Language and Cognitive Processes*, *16*(5), 507–534.

Dahan, D., & Tanenhaus, M. (2005). Looking at the rope when looking for the snake: Conceptually mediated eye movements during spoken-word recognition. *Psychonomic Bulletin & Review*, *12*(3), 453–459.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research*, *24*(6), 409–436.

Ehrlich, S., & Rayner, K. (1981). Contextual effects on word perception and eye movements during reading. *Journal of Verbal Learning & Verbal Behavior*, *20*(6), 641–655.

Fox Tree, J. E. (1995). The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, *34*, 709–738.

Fox Tree, J. E. (2001). Listeners' uses of um and uh in speech comprehension. *Memory & Cognition*, *29*, 320–326.

Fox Tree, J. E., & Clark, H. H. (1997). Pronouncing "the" as "thee" to signal problems in speaking. *Cognition*, *62*, 151–167.

Howell, P., & Young, K. (1991). The use of prosody in highlighting alteration in repairs from unrestricted speech. *Quarterly Journal of Experimental Psychology*, *43(A)*, 733–758.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Lau, E. F., & Ferreira, F. (2005). Lingering effects of disfluent material on comprehension of garden path sentences. *Language and Cognitive Processes*, *20*(5), 633–666.

Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104.

Lickley, R. J. (1995). Missing disfluencies. In *Proceedings of international congress of phonetic sciences* (Vol. 4, pp. 192–195).

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous speech. *Word*, *15*, 19–44.

Merlo, S., & Mansur, L. L. (2004). Descriptive discourse: Topic familiarity and disfluencies. *Journal of Communication Disorders*, *37*, 489–503.

O'Connell, D. C., & Kowal, S. (2005). Uh and um revisited: Are they interjections for signaling delay? *Journal of Psycholinguistic Research*, *34*, 555–576.

Pickering, M. J., & Garrod, S. (2007). Do people use language production to make predictions during comprehension? *Trends in Cognitive Sciences*, *11*(3), 105–110.

R Development Core Team. (2008). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org`

Schachter, S., Christenfeld, N., Ravina, B., & Bilous, F. (1991). Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology*, *60*, 362–367.

Schachter, S., Rauscher, F., Christenfeld, N., & Tyson Crone, K. (1994). The vocabularies of Academia. *Psychological Science*, *5*, 37–41.

Schnadt, M. J., & Corley, M. (2006). The influence of lexical, conceptual and planning based factors on disfluency production. In *Proceedings of the twenty-eighth meeting of the cognitive science society.* Vancouver, Canada.

Schwanenflugel, P. J., & Shoben, E. J. (1985). The influence of sentence constraint on

the scope of facilitation for upcoming words. *Journal of Memory and Language*, *24*(2), 232–252.

Shriberg, E. S. (1996). Disfluencies in Switchboard. In *Proceedings of the international conference on spoken language processing, Addendum* (pp. 11–14). Philadelphia, PA.

Shriberg, E. S. (2002). To errrris human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, *31*(01), 153–169.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Watanabe, M., Hirose, K., Den, Y., & Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, *50*(2), 81–94.

Appendix

Experimental Materials

Short Pause/No Retrace

Click on the red stripy uh ... spotty square
Click on the yellow spotty uh ... stripy circle
Click on the green stripy uh ... spotty star
Click on the blue spotty uh ... stripy triangle
Click on the yellow stripy uh ... spotty circle

Short Pause/Retrace

Click on the red stripy uh ... red spotty star
Click on the yellow spotty uh ... yellow stripy triangle
Click on the green stripy uh ... green stripy square
Click on the blue spotty uh ... blue spotty circle
Click on the green spotty uh ... green stripy triangle

Long Pause/No Retrace

Click on the blue stripy uh ... spotty square
Click on the green spotty uh ... stripy circle
Click on the yellow stripy uh ... spotty star
Click on the red spotty uh ... stripy triangle
Click on the red stripy uh ... spotty square

Long Pause/Retrace

Click on the blue stripy uh ... blue spotty star
Click on the green spotty uh ... green stripy triangle
Click on the yellow stripy uh ... yellow spotty square
Click on the red spotty uh ... red stripy circle
Click on the blue spotty uh ... blue stripy star