Hesitation in speech can... um... help a listener understand

Martin Corley (`Martin.Corley@ed.ac.uk`)
School of Philosophy, Psychology and Language Sciences
and Human Communication Research Centre
University of Edinburgh, Edinburgh EH8 9JZ, UK

Robert J Hartsuiker (`Rob.Hartsuiker@rug.ac.be`)
Department of Experimental Psychology
Ghent University, 9000 Ghent, Belgium

Abstract

This paper investigates the effect of disfluencies on listeners' on-line processing of speech. More specifically, it tests the hypothesis that filled pauses like *um*, which tend to occur before words that are low in accessibility, act as a signal to the listener that a relatively inaccessible word is about to be produced.

Two experiments are reported, in which participants followed recorded instructions to press buttons corresponding to images on a computer screen. In 50% of trials, the spoken name of the image was preceded by *um*. In experiment 1, the *intrinsic* accessibility of the target items was manipulated (by means of lexical frequency); in experiment 2, the *extrinsic* (visual) accessibility varied. Both experiments demonstrated that participants were quicker to respond when a target was preceded by *um*, regardless of whether the item referred to was difficult to access or not. In addition, in experiment 2 there was a weak interaction between accessibility and presence or absence of an *um*. We present the data here as early evidence that listeners can benefit from disfluencies in others' speech, and outline some methodological and theoretical considerations and further experiments.

By far the most common kind of language in use is conversation (Clark & Wilkes-Gibbs, 1987). In conversation, utterances are produced spontaneously. That is, they are "conceived and composed by their speakers even as they are spoken" (Mehta & Cutler, 1988, p. 136). A consequence of this is that spontaneous speech contains disfluencies. These are generally defined as "phenomena that interrupt the flow of speech and do not add propositional content to an utterance" (Fox Tree, 1995, p. 709). They include pauses, interruptions (midphrase or midword), repeated words and phrases, restarted sentences, words with elongated pronunciations, such as *the* pronounced /ðiː/ and *a* as /ɛiː/, and filled pauses such as *uh* and *um*. Such disruptions are very frequent: averaging across a number of studies, and excluding silent hesitations, Fox Tree (1995) estimated that the rate of disfluencies in spontaneous speech is about 6 words per 100 (see also Bortfield et al., 2001).

Despite the many disfluencies that occur in spontaneous speech, most studies of the comprehension of spoken language have used idealised, fluent utterances. This owes much to the commonly held view that disfluencies are noise and present obstacles to comprehension (Brennan & Schober, 2001, p. 275). However, some researchers have argued that disfluencies do not constitute "noise" at all, but are actually informative to listeners: they may provide information about the state of speakers' production systems. Specifically, certain disfluencies signal to listeners that speakers are experiencing production difficulty. Difficulty can occur at any stage of the process—during planning, lexical retrieval, or the articulation of a speech plan—and it has been argued that different types of disfluency signal different kinds of problems (e.g., Bortfield et al., 2001).

To date, much of the evidence supporting this account of conversational disfluencies has come from corpus studies of filled pauses such as *uh*, *um*, *the* as /ðiː/, and *oh* (e.g., Clark & Fox Tree, 2002; Fox Tree & Clark, 1997; Fox Tree & Schrock, 1999), and from experimental evidence gathered from speakers. For example, when asked to answer general knowledge questions, speakers tend to produce more *uh*s and *um*s before answers they are unsure of (Brennan & Williams, 1995; Smith & Clark, 1993). Moreover, *uh* appears to signal a shorter upcoming pause (and by inference, a less severe retrieval problem) than *um* (Smith & Clark, 1993), a finding borne out by corpus analyses (Clark & Fox Tree, 2002).

A number of studies examine in more detail the circumstances that might lead to a problem with retrieval. For example, unpredictable lexical items are preceded by hesitations more often than those that are predictable (Beattie & Butterworth, 1979). There is also a well-established correlation between disfluency and lexical frequency. For example, Maclay and Osgood (1959) examined a sample of spontaneous speech and found that "pauses filled with *er* and the like" were more likely to occur before open-class than (high frequency) closed-class words; Levelt (1983) showed that the frequency of colour names correlated negatively with the probability that these would be preceded by filled pauses.

However, findings concerning production do not necessarily imply that disfluencies are somehow *designed* to inform listeners about the states of speakers' production systems: they could simply be a by-product of the speech production process. Moreover, and of direct relevance to the current paper, they provide no evidence that listeners can or do *exploit* the information provided by disfluencies. For evidence of this kind, we turn to studies in which the focus is on the listener rather than the speaker.

Much of the reported evidence for listener sensitivity to disfluency comes from studies in which listeners are asked to compare or rate utterances. For example, Brennan and Williams (1995) presented participants with recorded answers to general knowledge questions, and asked them to estimate "how likely it was that the speaker knew the correct answer" (p. 389). Ratings were negatively affected by pauses before responses (as well as by the length of these pauses); but additionally, answers preceded by *uh* or *um* were judged less likely to be correct than answers preceded by unfilled (silent) pauses of the same lengths. Howell and Young (1991) found that listeners rated utterances including repairs as more comprehensible when those repairs were preceded by pauses. These studies, however, use off-line tasks (i.e., they measure comprehension after processing is complete). An important assumption underlying comprehension research is that comprehension takes place on-line (in "real time": Marslen-Wilson & Tyler, 1980). Assuming that disfluencies convey information, rather than noise, what we ultimately want to know is whether listeners can benefit from that information *as they comprehend* a given utterance.

Evidence that some disfluencies can immediately facilitate the comprehension of words that follow them comes from a study by Fox Tree (2001). Listeners identified words from recordings of speech with spontaneous *uh*s either present or digitally excised. Target words were recognised faster with the *uh*s present. Fox Tree concluded that *uh*s heighten attention to the speech that is to follow (cf. Fox Tree & Schrock, 1999, for *oh*). Similarly, Brennan and Schober (2001) found that compared to fluent controls, between-word interruptions (*yellow- orange*), and mid-word interruptions with or without fillers (*yell- uh- orange*, *yell- orange*) led to quicker identification of the "correct" (repair) word. The quickest identifications were in cases where the interruption included a filler. These studies however have a number of shortcomings. Fox Tree's materials may have had more natural prosodies with the *uh*s present; and the experimental task (word identification) is still some way from natural, on-line, language comprehension. Brennan and Schober (2001) did use a more natural task (following instructions referring to objects) but in their study the interruption itself reduced the potential number of referents. Participants in their study viewed a display with two objects, one of which was the target. When the naming of one was interrupted, it was immediately clear that the other was the target, thus enabling participants to respond fast. This artefact, rather than the repairs *per se*, may account for their findings.

At best, the on-line studies above provide only weak evidence for the proposition that listeners can exploit such information as might be conveyed by disfluencies. This is because the information content in each study is low: if disfluencies are supposed to signal problems in access for the speaker, the listener needs to be able to judge which parts of an utterance (here, references to objects or concepts) are likely to be difficult. In the above

studies, accessibility is not manipulated: all referents are equally accessible or inaccessible.

To date, two studies have directly manipulated accessibility in tests of listeners' sensitivity to disfluency. Barr (2001) and Arnold, Fagnano, and Tanenhaus (2003) manipulated target words in terms of accessibility with respect to the discourse model. It is a common finding that newly introduced items are harder to access than (recently mentioned) items already in the discourse (e.g., Arnold, Wasow, Ginstrom, & Losongco, 2000), because new information has a lower expectancy. Barr (2001) presented listeners with sentences describing abstract shapes that were either familiar or new to them. Their task was to point to the shape that matched the description they were hearing. When the shapes constituted new information, listeners' responses were faster when descriptions were preceded by an *um* than when they were preceded by random noise. Arnold et al. (2003) conducted a study in which participants' eye movements were recorded while they viewed a series of displays of four objects, two of which began with the same phonological segments (e.g., *candle* and *camel*). Participants were instructed to move one of these two competitor objects, which had either been established as discourse-new or discourse-given; a proportion of instructions contained disfluencies. Arnold et al. found that, regardless of the content of the instructions, more initial fixations were made on the discourse-given competitor when the instruction was fluent, whereas a disfluent instruction led to more initial fixations on the discourse-new object.

Although the studies by Arnold et al. (2003) and Barr (2001) are highly suggestive, they still leave some questions open. The first is that of the *types* of accessibility information which listeners can make use of when they encounter disfluent speech. It is well established that disfluency can result from language-internal, or *intrinsic*, accessibility difficulties (for example, speakers are more likely to be disfluent when naming low-frequency colours; Levelt, 1983). Equally, when accessibility is *extrinsic*, that is, manipulated independently of language, speakers are likely to be disfluent. For example, speakers are likely to use more filled pauses when describing ambiguous pictures (Siegman & Pope, 1966). The distinction between the two types of accessibility is important, because it is only in the instrinsic case that listeners must effectively be able to model the production process.

The second question is that of what is is about a disfluent utterance that cues listeners to use the information. Barr (2001) and Arnold et al. (2003) used naturalistic recordings of spontaneous speech as auditory stimuli in their experiments. Although there are clear advantages to this approach in terms of ecological validity, it still leaves open the question of whether something other than a filled pause (such as, say, the prosody of an entire utterance, as suggested by Arnold et al., 2003) is acting to cue the listener to pay attention to a given object.

In the experiments reported below, we aim to explore listeners' sensitivity to intrinsic (experiment 1) and extrinsic (2) accessibility information, in the face

of speaker disfluency, by manipulating the lexical frequency (1) and visual accessibility (2) of items referred to in spoken instructions. In contrast to the studies reported above, we will use digitally edited recordings, with the *only* difference between fluent and disfluent utterances being the presence or absence of an *um*. If listeners are sensitive to disfluency and can make use of the appropriate type of accessibility information, we expect them, in each experiment, to be faster to respond to *fluent* instructions mentioning *accessible* items, but to *disfluent* instructions referring to *inaccessible* objects.

Experiment 1

Experiment 1 was designed to investigate whether listeners were sensitive filled pauses preceding references to items that differed in intrinsic accessibility (exemplified in this case by lexical frequency). The experiment made use of a reaction time paradigm in which participants were presented with pairs of pictures of everyday objects. In each case, one picture corresponded to a high-frequency lexical item, and the other to a low-frequency item. Upon hearing an auditory instruction naming one of the objects, participants had to respond by pressing one of two buttons (corresponding to the left-had or right-hand object). 50% of the instructions included a filled pause (*um*) just before the object was named. The two factors manipulated orthogonally in a within-subjects design were lexical frequency of the named item, and fluency of instruction. Recordings were made of times (relative to the onset of the target item name) taken for accurate responses to the instructions.

Method

Participants 32 students at the University of Edinburgh volunteered to take part in the experiment, which lasted approximately 5 minutes. None reported having hearing difficulties.

Materials The experimental materials consisted of auditory and visual stimuli. The latter were pairs of pictures with high- and low-frequency names. The pictures were coloured versions of a subset of the standardised picture set used by Snodgrass and Vanderwart (1980), and were normed for frequency, visual complexity, and familiarity (Rossion & Portois, 2001). 16 HF pictures (mean frequency 300 occurrences per million; range 153–796) and 16 LF pictures (mean frequency 5.29; range 0.22–9.89) were paired four times (never in the same combinations), resulting in 64 picture pairs. Three pairs of mid-frequency items (lamp-cake, clock-knife, wheel-cow) were used for practice trials at the start of the experiment. No picture depicted an word that started with a vowel (because this would be preceded by /ði/ in an instruction), and no pair of pictures represented words that began with the same phoneme, or had semantic overlap. Each individual picture was on the left of the screen for two of the four times it appeared, and on the right for the remainder. It was a target twice: once each for a fluent and disfluent instruction, once on each side of the screen.

now press the button for the   um...   broom, please

Figure 1: Example recording showing insertion of *um*

The auditory stimuli consisted of instructions to press a button corresponding to a particular picture. They were always of the form *now press the button for the [um] <target>, please*. Similar materials have been successfully used in eyetracking experiments (cf. Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995). Using invariant instructions has the additional advantage of controlling the syntactic and semantic context in which the target words occur. There were two versions of each instruction for each picture in the set of 32: a fluent one, and a disfluent one in which an *um* had been artificially inserted. This resulted in 64 utterances in total.

To make the recordings, a native speaker of English read a list with each target item embedded in the template instruction sentence (see above). After the recording, all target items were removed from their original contexts, together with the word *please* which followed them, and spliced into one version of the carrier sentence which had not originally included any of the target items. This resulted in a set of 32 fluent instructions, for each of which the target word onset was exactly 1218 ms. after the utterance onset. To make the disfluent instructions, the speaker was asked to read a number of instructions referring to low-frequency items, inserting an *um* "as naturally as possible". The single *um* that sounded best was selected, and was spliced into each of the fluent instructions immediately before the target word, to create a matched set of disfluent utterances. Since the *um* was 1078 ms. long, the target onset in disfluent instructions occurred 2296 ms. after the utterance onset.

Finally, each recording was converted to a 16-bit 22050Hz stereo WAV file, for use with E-Prime experimental software. An example plot of a disfluent instruction can be seen in figure 1.

Procedure Participants were tested individually in a quiet room. They were informed that they were participating in an experiment on sentence comprehension, and that they would be listening to a series of recordings of a speaker giving instructions as fast as possible. The aim of the study was purportedly to establish how easy it is to follow instructions given in stressful situations. This minor deception was necessary to justify the disfluencies in the study.[1] Participants were instructed that they would

---

[1] Although the number of disfluencies in the experiment (32 in 5184 words) was well below the 6% cited by Fox Tree (1995) they were nonetheless highly salient (in fact many disfluencies go undetected: Lickley, 1995).

Table 1: Experiment 1: Mean correct RT, relative to target onset (s.e. in brackets)

Instructions	Target Frequency	
	high	low
- um	625 (25.8)	650 (27.3)
+ um	558 (22.6)	604 (26.1)

Table 2: Experiment 2: Mean correct RT, relative to target onset (s.e. in brackets)

Instructions	Target Type	
	clear	blurred
- um	665 (50.1)	731 (49.6)
+ um	612 (45.6)	653 (41.9)

be presented with a series of displays of pairs of pictures. Each picture pair would be accompanied by instructions to press the button corresponding to a given object. They had a 5-button button-box in front of them: if the picture referred to was on the right, they were to press the rightmost button; if on the left, the leftmost button. It was stressed that they should respond as quickly as possible, without losing accuracy.

Prior to the experiment, three practice items allowed the participants to familiarise themselves with the procedure, and to adjust the volume on the headphones they wore to hear the instructions. The practice session was identical to the experimental session in all respects bar one; the 3 practice items were always presented in a fixed sequence, whereas the presentation order of the 64 experimental items was randomised.

In the practice sessions as well as in the experiment proper, each display of a picture pair was preceded by a "+", which remained visible for 200 ms., to signal that a new pair of pictures was about to come up. The pictures followed this display immediately. At the same time as the pictures appeared, the corresponding instruction was played. Each instruction was played in full, regardless of whether or not a button had been pressed before it ended. The instructions always finished before the pictures were removed, 4 seconds after onset. Once each trial had finished, the screen was blanked, and the next trial began with a "+" after a 250 ms. pause. The time between the onset of the instruction and the corresponding button press was recorded for each correct response.

Results

Prior to analysis, all reaction times were converted to times relative to the target onset: that is, 1218 ms. was subtracted from the time for each response recorded to a fluent instruction, and 2296 ms. from responses to disfluent instructions. The resulting RTs for correct responses were analysed by participants (F1) and by items (F2).

Table 1 shows the mean reaction times by participants. There was a significant main effect of disfluency: participants were faster to respond when the instructions included an *um* (581 vs. 638 ms.; $F1(1, 30) = 35.21$, $p < .001$; $F2(1, 30) = 51.81$, $p < .001$). The effect of frequency was only significant by participants ($F1(1, 30) = 19.95$, $p < .001$; $F2(1, 30) = 3.45$, $p = .073$).[2] Unfortunately, there is no sign of an appropriate interaction in table 1: if anything, high-frequency items are more advantaged by disfluent instructions than

low-frequency items (however, this interaction is not significant).

Discussion

One explanation of these findings is that *um* does not convey useful information to listeners: instead, it may simply focus attention, allowing for more rapid processing (cf. Fox Tree, 2001). However, it is also possible that it is the *type* of accessibility that is at fault in this experiment. That is, listeners may simply not be sensitive to the interaction of disfluency and intrinsic features such as lexical fluency. Accordingly, experiment 2 is a replication of experiment 1 in which the *extrinsic* accessibility of the stimuli is manipulated.

Experiment 2

In experiment 2, all of the pictures from experiment 1 corresponding to low frequency names were blurred, over a radius of 15 pixels, using an image editor. Thus the low-frequency items from experiment were now extrinsically difficult to access: the primary difficulty with retrieval was external to the language system.

22 volunteer participants, all students at Edinburgh University, took part in the experiment. In all respects other than the images used (design, procedure, analysis) it was identical to experiment 1.

Results

Table 2 shows the mean reaction times by participants. Once again, there was a main effect of disfluency, with participants faster to respond after *um* (632 vs. 698 ms.; $F1(1, 21) = 45.25$, $p < .001$; $F2(1, 30) = 66.95$, $p < .001$). In this experiment, the effect of accessibility was also clearly significant: there were longer responses to blurred items (692 vs. 638 ms.; $F1(1, 21) = 59.46$, $p < .001$; $F2(1, 30) = 7.02$, $p = .013$). Allowing for the *um* advantage, the mean reaction times are in the predicted direction: blurred items are speeded up more by *um* (78 ms.) than are clear items (53 ms.). Although the differences are not large, this interaction is marginal by items ($F1(1, 21) = 2.47$, $p = .131$; $F2(1, 30) = 4.08$, $p = .052$).

---

[2]The lack of a by-items effect for frequency suggests problems with some of the items, rather than an insensitivity to frequency on the part of participants. A post-hoc analysis, in which the four slowest HF items and the four fastest LF items were removed, suggests that this is the case ($F1(1, 30) = 63.06$, $p < .001$; $F2(1, 22) = 22.27$, $p < .001$).

Discussion

Once again the clearest effect in experiment 2 is the "*um* advantage" (we return to this point in the main discussion below). However, there are some signs of an interaction between accessibility and fluency: a tentative conclusion would be that *um* may be more than a content-free focusing device; a more concrete conclusion is that further investigation is certainly warranted.

General Discussion

Although the experiments reported here do not provide conclusive evidence for listeners' ability to interpret *um* as indicating speakers' difficulties with speech, the results are suggestive and indicate a potentially productive line of research. In particular, the "*um* advantage" observed in both experiments may be masking more subtle effects. Although a number of researchers have suggested that disfluency may play a focusing role (e.g., Brennan & Schober, 2001; Fox Tree, 1995, 2001; Fox Tree & Clark, 1997; Fox Tree & Schrock, 1999), there is a possibility that the faster reaction times after *um* in these experiments are an artefactual result of entropy (assuming that participants in this experiment were able to gain an implicit understanding of the experimental design[3]) : consider having heard the words *now press the button for the...* In cases where there is no *um*, there are still 3 possibilities at this point: *um*, or one of two target words. On the other hand, an *um* signals that the next constituent *must* be a target word, and participants may be able to use this information strategically to prepare a response. A third experiment (currently underway) addresses this problem by contrasting *now press the button for the um...* with *now press the um button for the...*: in these cases, participants always know immediately prior to the target word that the next word they encounter must be a target. If the artefactual account outlined here is correct, the *um* advantage should disappear.

A number of claims have been made in the literature concerning the function of disfluency in conversational speech, both from the perspective of the speaker (who according to some accounts is "conveying a message" via disfluency) and from that of the listener, who may be able to make use of disfluency to glean information about the speaker's current status. Unfortunately, many of these claims are poorly supported by the existing evidence. In this paper, we have attempted to demonstrate the beginnings of a methodical approach to the claims that have been made, and highlighted an important distinction between *types* of accessibility in terms of what the listener needs to know about the linguistic difficulty of retrieving an object's name.

References

Arnold, J. E., Fagnano, M., & Tanenhaus, M. K. (2003). Disfluencies signal thee, um, new information. *Journal of Psycholinguistic Research*, *33*, 25–36.

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

Barr, D. J. (2001). Trouble in mind: Paralinguistic indices of effort and uncertainty in communication. In S. Santi, I. Guaitella, C. Cave, & G. Konopczynski (Eds.), *Oralité et gestualité, communication multimodal, intéraction.* Paris, France: L'Harmattan.

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

Bortfield, H., Leon, S. D., Bloom, J. E., Schober, M. F., & Brennan, S. E. (2001). Disfluency rates in spontaneous speech: Effects of age, relationship, topic, role, and gender. *Language and Speech*, *44*, 123–147.

Brennan, S. E., & Schober, M. F. (2001). How listeners compensate for disfluencies in spontaneous speech. *Journal of Memory and Language*, *44*, 274–296.

Brennan, S. E., & Williams, M. (1995). The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, *34*, 383–398.

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

Clark, H. H., & Wilkes-Gibbs, D. (1987). Referring as a collaborative process. *Cognition*, *22*, 1–39.

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 and 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.

Fox Tree, J. E., & Schrock, J. C. (1999). Discourse markers in spontaneous speech: Oh what a difference an oh makes. *Journal of Memory and Language*, *40*, 280–295.

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.

---

[3]This assumption is based on the fact that in debriefing, some participants stated that they had become aware that the recorded instructions often included an *um* before the target name.

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

Lickley, R. J. (1995). Missing disfluencies. In *Proceedings of the international congress of phonetic sciences* Vol. 4. Stockholm, Sweden.

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

Marslen-Wilson, W., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, *8*, 1–71.

Mehta, G., & Cutler, A. (1988). Detection of target phonemes in spontaneous and read speech. *Language and Speech*, *31*, 135–156.

Rossion, B., & Portois, G. (2001, May). *Revisiting Snodgrass and Vanderwart's object database: Color and texture improve object recognition.* (Paper presented at the 1st Vision Science Conference, Sarasota)

Siegman, A. W., & Pope, B. (1966). Ambiguity and vebal fluency in the TAT. *Journal of Consulting Psychology*, *30*, 239–245.

Smith, V. L., & Clark, H. H. (1993). On the course of answering questions. *Journal of Memory and Language*, *32*, 25–38.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*.

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*, 1632–1634.