Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Cognition 105 (2007) 658–668
COGNITION
www.elsevier.com/locate/COGNIT

Brief article

It's the way that you, er, say it: Hesitations in speech affect language comprehension

Martin Corley [a], Lucy J. MacGregor [a,*], David I. Donaldson [b]

[a] *School of Philosophy, Psychology, and Language Sciences, University of Edinburgh, Edinburgh EH8 9JZ, UK*
[b] *Department of Psychology, University of Stirling, FK9 4LA, UK*

Received 12 July 2006; revised 29 September 2006; accepted 29 October 2006

Abstract

Everyday speech is littered with disfluency, often correlated with the production of less predictable words (e.g., Beattie & Butterworth [Beattie, G., & Butterworth, B. (1979). Contextual probability and word frequency as determinants of pauses in spontaneous speech. *Language and Speech, 22*, 201–211.]). But what are the effects of disfluency on listeners? In an ERP experiment which compared fluent to disfluent utterances, we established an N400 effect for unpredictable compared to predictable words. This effect, reflecting the difference in ease of integrating words into their contexts, was reduced in cases where the target words were preceded by a hesitation marked by the word *er*. Moreover, a subsequent recognition memory test showed that words preceded by disfluency were more likely to be remembered. The study demonstrates that hesitation affects the way in which listeners process spoken language, and that these changes are associated with longer-term consequences for the representation of the message.

Keywords: Language comprehension; Disfluency; Speech; ERPs

* Corresponding author. Tel.: +44 131 651 1304; fax: +44 131 650 3461.
E-mail address: Lucy.MacGregor@ed.ac.uk (L.J. MacGregor).

0010-0277/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cognition.2006.10.010

1. Introduction

Approximately 6 in every 100 words are affected by disfluency, including repetitions, corrections, and hesitations such as the fillers *um* and *er* (Fox Tree, 1995). Moreover, the distribution of disfluency is not arbitrary. For example, fillers tend to occur before low frequency and unpredictable words (Beattie & Butterworth, 1979; Levelt, 1983; Schnadt & Corley, 2006), in circumstances where the speaker is faced with multiple semantic or syntactic possibilities (Schachter, Christenfeld, Ravina, & Bilous, 1991), as well as in cases where other types of uncertainty occur (Brennan & Williams, 1995). But what are the effects of hesitations on listeners and on language comprehension?

Although the majority of psycholinguistic research on speech comprehension has been conducted using idealised, fluent utterances, a number of corpus analyses and behavioural studies suggest that disfluency can affect listeners. Longer-term consequences of disfluency include speakers being rated as less likely to know answers to general knowledge questions when their answers are preceded by hesitations (Brennan & Williams, 1995), suggesting that listeners are sensitive to the uncertainty conveyed by hesitations at a metacognitive level. Offline questionnaire studies additionally reveal that hesitations can influence grammaticality ratings for garden path sentences, reflecting probable differences in the ways in which they have been comprehended (Bailey & Ferreira, 2003).

Investigations of the shorter-term effects of disfluency show that listeners are faster at a word monitoring task when words are preceded by a hesitation (Fox Tree, 2001) and from this it has been argued that hesitations heighten listeners' immediate attention to upcoming speech. Work by Arnold and colleagues (Arnold, Tanenhaus, Altmann, & Fagnano, 2004) attempts to refine an account of how listeners respond to disfluency in real time. Using a visual world paradigm (Tanenhaus, Spivey-Knowlton, Eberhard, & Sedivy, 1995), participants' eye movements to depictions of objects on a computer screen were monitored as they responded to auditory instructions to move the objects with a mouse. The presence of a disfluency (*thee uh*) before the target object increased the probability of an initial eye movement to an object that had not been previously mentioned; in contrast, when the instructions were fluent, participants were more likely to look first at a previously mentioned object. Arnold et al.'s (2004) findings suggest that listeners are sensitive to the fact that speakers find it more difficult to retrieve the names of items they have not mentioned before (Arnold, Wasow, Ginstrom, & Losongco, 2000) and can predict that these items are more likely to be mentioned following disfluency. However, there are at least two limitations of the Arnold et al. (2004) study. First, the effects of disfluency may be driven by the nature of the task. In natural dialogue, it is rare for listeners to be presented *a priori* with a limited set of images which provide potential sentence completions (although see Dahan, Magnuson, & Tanenhaus, 2001, for evidence that non-presented items can affect eye-movements). Second, the study does not address possible longer-term consequences of disfluency. Therefore, although the results suggest that listeners can strategically profit from disfluency in a constrained task-driven situation, the question of whether and how disfluency affects

listeners on-line and in the longer term, under more natural circumstances, remains unanswered.

Our study addresses both of these issues, using Event-Related Potentials (ERPs) to provide a real-time measure of the processing of disfluent speech, and a surprise recognition memory test to assess the longer-term consequences of disfluency on language representation.

ERPs – neural activity recorded at the scalp, time locked to the onset of a cognitive event of interest and averaged over multiple events – are ideal for investigating the functional and neural basis of spoken language comprehension. They have two particular benefits over eye movement methods. First, there is no need for a contextually relevant visual presentation (with its attendant constraints), and second, participants need not perform any other task other than listen to the experimental stimuli. This means that ERPs provide an ideal means to investigate how listeners process disfluent speech in a situation which is a close analogue to everyday language comprehension.

We focused on the N400, an ERP component associated with the meaningful processing of language (Kutas & Hillyard, 1980, 1984). During comprehension, each word must be integrated with its linguistic context, from which it can often be predicted. Where integration is difficult (for example because a word is not predictable), a negative change in voltages recorded at the scalp relative to more easily integrated words is observed. This difference, the N400 effect, peaks at around 400 ms after word onset, maximally over central and centro-parietal regions.

Because disfluency tends to precede less predictable items in speech (Beattie & Butterworth, 1979; Levelt, 1983; Schnadt & Corley, 2006), we focused on listeners' ability to integrate predictable and unpredictable target words into their preceding contexts. If listeners interpret hesitation as a signal that the following words may not follow from the preceding context, the presence of hesitations before target words should reduce the N400 difference between predictable and unpredictable words. Changes in the N400, indicating differences in the processing of the input, may result in changes to the representation of the message, particularly of the words immediately following the disfluency. An effect in memory for these words would provide evidence for this, as well as a longer-term correlate of any effects observed in the ERP record at the time the utterances were heard.

2. Method

2.1. Materials

Auditory materials were created from 80 pairs of sentence frames, together with corresponding pairs of utterance-final target words, which were the most predictable ending for one sentence frame (mean cloze probability: .84) and an unpredictable ending for the other (0). Predictability was determined using a cloze probability pre-test. Table 1 shows an example material set. Double-counterbalancing ensured that each target word and each sentence frame contributed equally to each of the

Table 1
Example stimulus set comprising two highly constraining sentence frames, crossed with two target words which were predictable or unpredictable in context

Predictable	Everyone's got bad habits and mine is biting my	[*er*]	nails
	That drink's too hot; I have just burnt my	[*er*]	tongue
Unpredictable	Everyone's got bad habits and mine is biting my	[*er*]	tongue
	That drink's too hot; I have just burnt my	[*er*]	nails

Recorded utterances were either fluent or disfluent (containing the filler *er*, indicated in square brackets).

conditions obtained from crossing disfluency with predictability, and that no participant heard any of the sentence frames or target words twice.

Fluent and disfluent versions of the sentence frames were recorded at a natural speaking rate. In each case, the target word was replaced with a 'pseudo-target' (e.g., *pen*) so that actual targets were not predictable from phonotactic cues in the frames. In disfluent versions, the pseudo-target was preceded by an *er* (pronounced [ɜː]) with prolongations of the previous word (e.g., *thee* [ðiː]), and included prosodic changes where natural for the speaker. Finally, identical recordings of the target words were spliced onto the recorded frames in place of the pseudo-targets. This ensured that any observed ERP differences between conditions would be directly attributable to the contexts, rather than to differences between the recordings of the targets themselves. In each of four versions of the experiment, 80 of the resulting recordings were presented in disfluent form, and 80 were fluent. Recordings of 80 unrelated filler sentences, including some with less predictable words either mid-utterance or at the end of the utterance, were also added to each version. Half of the fillers included disfluencies of various types.

2.2. Participants

Twelve native British English speakers (6 male; mean age 23; range 16–35; all right-handed) with no known hearing or neurological impairment participated for financial compensation. Informed consent was obtained in accordance with the University of Stirling Psychology Ethics Committee guidelines.

2.3. Procedure

There were two parts to the experiment. In the first part, participants were told that they would hear a series of utterances which were re-recorded extracts from previously recorded conversations, and that they should listen for understanding, just as they would in a natural situation. No other task was imposed. One hundred and sixty experimental utterances were presented auditorily, interspersed with fillers. Recordings were presented in two blocks lasting approximately 15 min each, separated by a break of a few minutes. The start of each recording was signalled visually by a fixation cross, used to discourage eye movements.

EEG was recorded from 61 scalp sites using a left mastoid reference, and re-referenced to average mastoid recordings off-line. Electro-oculograms were recorded

to monitor for vertical and horizontal eye movements. Electrode impedances were kept below 5 kΩ. The analogue recordings were amplified (band pass filter 0.01–40 Hz), and continuously digitised (16 bit) at a sampling frequency of 200 Hz.

Before off-line averaging, the continuous EEG files for each participant were screened, resulting in a loss of 24.8% of ERP trials due to artefacts, with little variability across conditions. The effect of blink artefacts was minimised by estimating and correcting their contribution to the ERP waveforms (Rugg, Mark, Gilchrist, & Roberts, 1997). Average ERPs (epoch length 1350 ms, pre target baseline 150 ms) time locked to the onsets of target words were formed for each participant (average 26 artefact-free trials by condition, minimum 16), and the waveforms were smoothed over 5 points.

In the second part of the experiment, the 160 utterance-final 'old' (previously heard) words were presented visually, interspersed with 160 frequency-matched 'new' foils, which had not been heard at any point in the first part of the experiment. Participants discriminated between old and new words as accurately as possible by pressing one of two response keys. The start of each presentation of a target word was signalled by the appearance of a fixation cross, which was replaced by the stimulus. After a 750 ms presentation, the screen was blanked for 1750 ms. Responses made later than this were not recorded.

3. Results

ERPs in response to predictable and unpredictable target words in fluent and disfluent utterances were quantified by measuring the mean amplitude within the standard N400 time window of 300–500 ms after word onset. All analyses made use of Greenhouse–Geisser corrections where appropriate, and are reported using corrected F values.

Figs. 1 and 2 show ERPs time locked to the utterance-final word onsets for fluent and disfluent utterances respectively, for midline (Fz, FCz, Cz, CPz, Pz) and grouped left- and right-hemisphere electrodes. Unpredictable words lead to greater negativity over the conventional N400 epoch of 300–500 ms. This negativity is broadly distributed over the scalp, but appears larger over central and midline locations, closely resembling effects shown in previous studies (Kutas & Hillyard, 1980, 1984; Van Berkum, Brown, & Hagoort, 1999; Van Petten, Coulson, Rubin, Plante, & Parks, 1999).

Because the pre-target baselines for fluent and disfluent materials were recorded from different points in the utterances (disfluent baselines are typically obtained mid-*er*), direct comparisons for targets in fluent vs. disfluent conditions could not be made: instead we used an interaction analysis to compare the size of the N400 predictability effect across conditions. In order to establish that this comparison was meaningful, we first ensured that there was no distributional difference between the N400s obtained in fluent and disfluent conditions. To do this, we calculated the mean voltage difference between ERPs for unpredictable and predictable targets over the 300–500 ms time window for each of the 61 electrodes, separately for fluent and disfluent utterances. ANOVA (factors of fluency and location) performed on

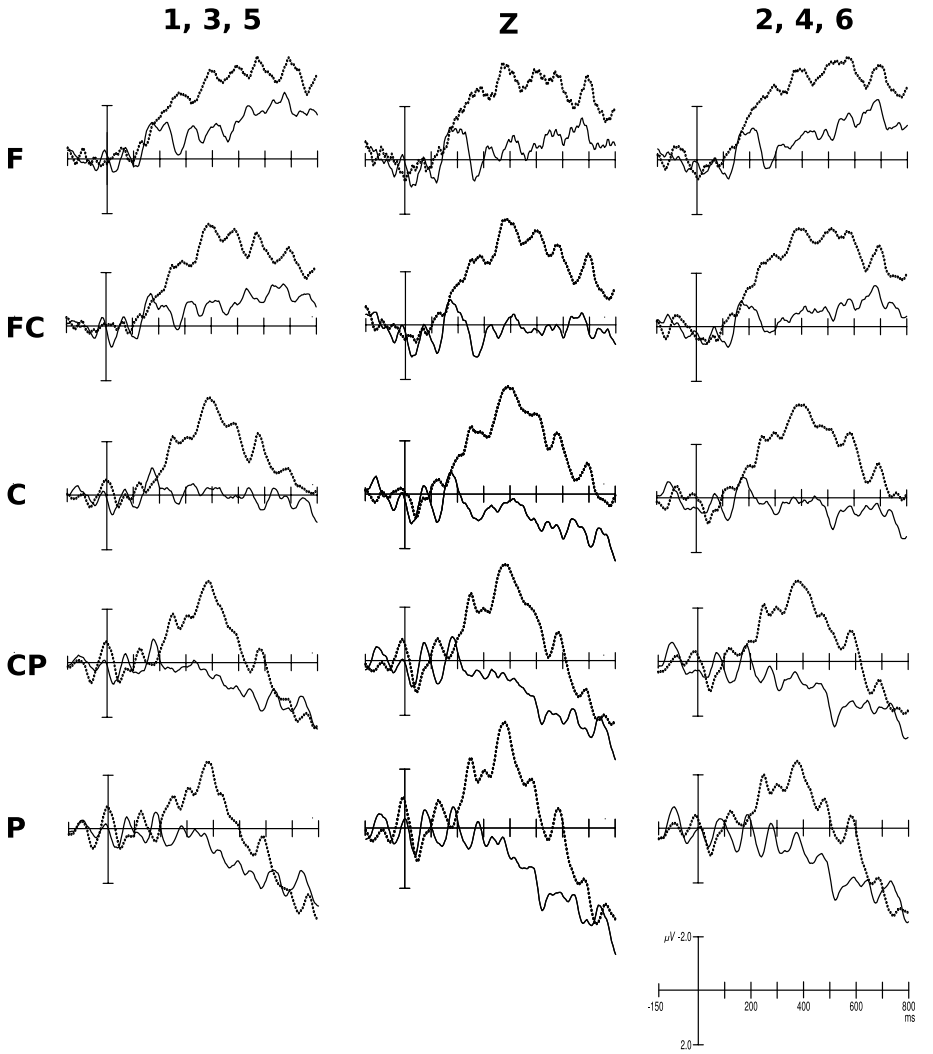

Fig. 1. ERPs for *fluent* utterances relative to predictable (solid lines) or unpredictable (dotted lines) target word onsets. The central column represents the midline sites (from top: frontal, fronto-central, central, centro-parietal, parietal); the left-hand and right-hand columns represent averages of three electrodes to the left or right of the midline, respectively.

these differences, after normalisation for amplitude differences (using the max/min method: McCarthy & Wood, 1985), reveals no effect of location [$F(60, 660) = 1.70$, $\epsilon = .046$, $\eta_p^2 = .134$, $p = .191$], nor of fluency [$F < 1$], nor any interaction between fluency and location [$F < 1$]. The lack of difference in scalp topographies between the fluent and disfluent conditions gives us no reason to suppose that different neural generators are responsible for the recorded effects of predictability.

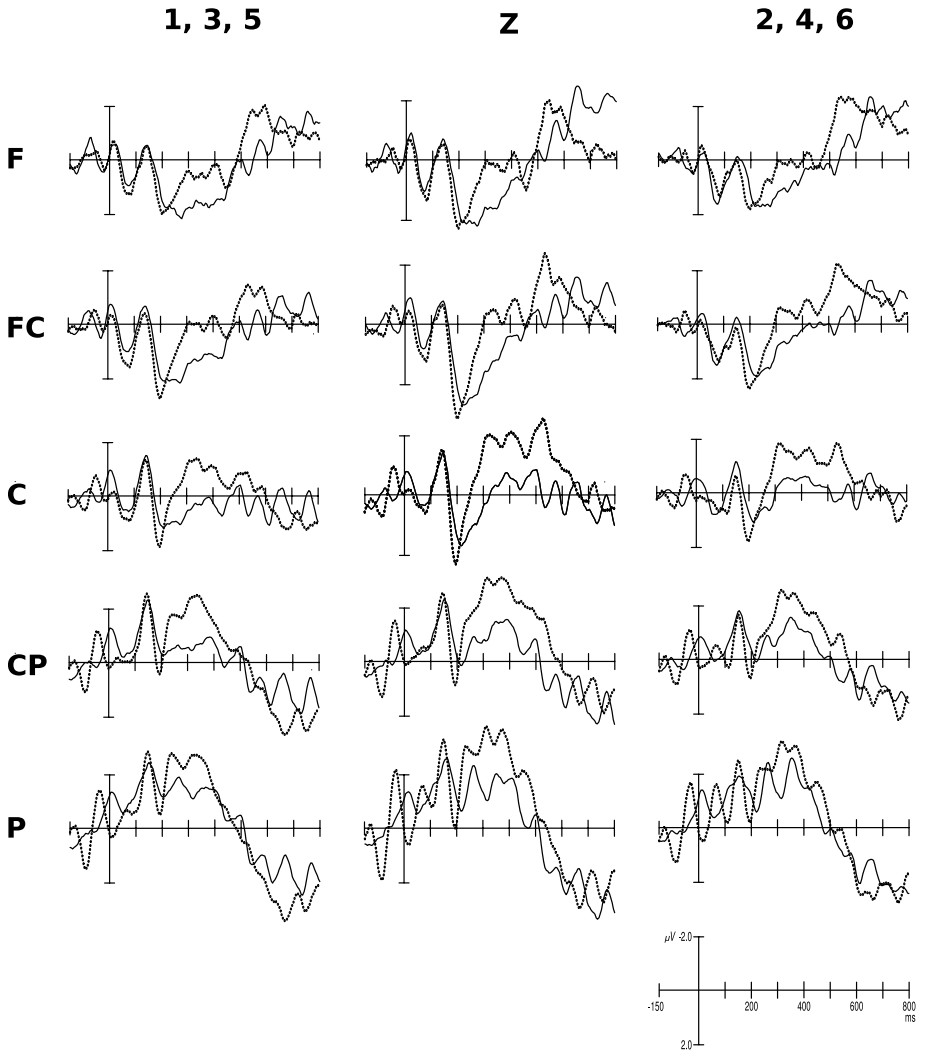

Fig. 2. ERPs for *disfluent* utterances relative to predictable (solid lines) or unpredictable (dotted lines) target word onsets. The central column represents the midline sites (from top: frontal, fronto-central, central, centro-parietal, parietal); the left-hand and right-hand columns represent averages of three electrodes to the left or right of the midline, respectively.

Two further analyses established that the distributions of the fluent and disfluent N400s were not lateralised (factors of predictability, location [F, FC, C, CP, P], hemisphere [L, R], and laterality [1, 2 vs. 3, 4 vs. 5, 6]). For fluent utterances, the analysis revealed a main effect of predictability [$F(1,11) = 43.93$, $\eta_p^2 = .800$, $p < .001$] and an interaction of predictability with laterality [$F(2,22) = 8.95$, $\epsilon = .550$, $\eta_p^2 = .448$, $p = .010$]. No other effect involving predictability was significant.

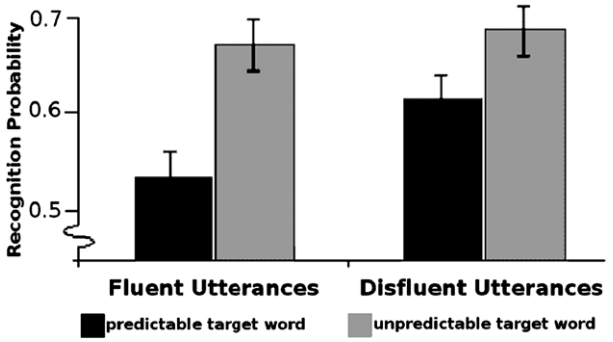

Fig. 3. Memory performance for utterance-final words which were originally predictable or unpredictable in their contexts, by utterance fluency (error bars represent one standard error of the mean).

For disfluent utterances, there was no effect of predictability [$F(1,11) = 2.96$, $\eta_p^2 = .212$, $p = .113$] and no other effect involving predictability reached significance. Since no effects involving hemisphere were found in either analysis, we concentrated on the midline electrodes (Fz, FCz, Cz, CPz, Pz) for the comparison of fluent with disfluent utterances.

An analysis of the midline electrodes (factors of fluency, predictability, location) demonstrated a main effect of predictability [$F(1,11) = 19.39$, $\eta_p^2 = .638$, $p = .001$] and an interaction of fluency with location [$F(4,44) = 13.79$, $\epsilon = .307$, $\eta_p^2 = .556$, $p = .002$], reflecting general frontal positivity relative to the baseline in the disfluent case. Importantly, fluency interacted with predictability [$F(1,11) = 7.93$, $\eta_p^2 = .419$, $p = .017$], establishing that the N400 effect for fluent items (3.14 μV) is reduced for disfluent items (1.19 μV).

As a final check, we performed an ANOVA for the midline N400 effects after normalisation (using the max/min method) to examine whether there were any distributional differences between fluent and disfluent conditions for this crucial interaction. There were no observable differences between fluent and disfluent items [for location: $F(4,44) = 1.34$, $\epsilon = .274$, $\eta_p^2 = .109$, $p = .274$; other Fs < 1].

Memory performance was quantified as the probability of correctly identifying old (previously heard) words. To control for differences in individual memory performance, we treated stimulus identity as a random factor.[1] Overall, 62% of the old words were correctly recognised (false alarm rate 24%). Fig. 3 shows the recognition probability of utterance-final words by fluency and predictability.

ANOVA (factors of fluency and predictability) reveals that words which were unpredictable utterance endings are more likely to be recognised than predictable

[1] Traditional adjustments for individual error-rates, such as d', are inappropriate here, since the properties of 'old' stimuli are determined by their context of occurrence and hence there are no comparable categories of 'new' stimuli. Using stimulus identity as a random factor ensures that per-participant biases to respond 'old' or 'new' are controlled for across the experiment.Twelve target words were inadvertently repeated in the experiment, resulting in 148 distinct targets. Analysis with data from the repeated targets removed did not affect the outcome.

words [69% vs. 58%: $F(1,147) = 23.48$, $\eta_p^2 = .138$, $p < .001$]. Importantly, disfluency also has a long-term effect: words which were preceded by hesitation are better recognised [66% vs. 62%: $F(1,147) = 4.31$, $\eta_p^2 = .029$, $p < .05$], primarily predictable words [62% vs. 55%: $F(1,147) = 4.73$, $\eta_p^2 = .031$, $p = .031$].

4. Discussion

In the presence of disfluency, the N400 effect, traditionally associated with the processing of less compared to more predictable words, was substantially reduced. Hesitation also had a longer-term effect: words following *er* were more likely to be recognised in a subsequent memory test. This suggests that these words have been processed differently as a consequence of hesitation. Since the N400 differences correspond to differences in memory performance, we can additionally conclude that the ERP differences are not due to contamination of the N400 waveform by spillover effects from the processing of the *er*.

Because predictability and ease of integration are often confounded, we are left with two possible accounts of the locus of the N400 attenuation. First, it may be because the *er* affects post-lexical factors, which operate once the target has been heard. Previous research has shown that the N400 is sensitive to differences in the semantic fit of words that do not differ in terms of predictability (Van Berkum, Zwitserlood, Hagoort, & Brown, 2003). We know from the speech production literature (e.g., Levelt, 1989) that fillers such as *er* often co-occur with other disfluent phenomena such as corrections. These are hypothesised to be more difficult to integrate syntactically and semantically, because some kind of revision must take place. A similar process could be responsible for post-*er* integration in the current experiment: hesitation could add to the difficulty with which both predictable and unpredictable words are integrated. Alternatively, the *er* may affect the comprehension system before the target is heard, effectively reducing the extent to which specific predictions are made, and therefore increasing the integration difficulty. In both cases, we might assume that predictable words would give rise to more negativity in disfluent compared to fluent contexts, as suggested by visual inspection of Figs. 1 and 2. Since limitations of the present design prevent a direct comparison from being made, it is particularly relevant that these views also predict that words following disfluency will be better remembered. Such a memory effect is demonstrated in this study, albeit with a small size because of the large number of other factors that are likely to influence the likelihood of later remembering a particular word heard among 240 recorded utterances.

Whatever the detailed mechanism, disfluency clearly affects the processing of language. But what is it about *er* that causes a processing change? One view is that there is nothing intrinsic to *er* that allows it to be understood as a disfluent signal. Instead, the N400 attenuation and subsequent effects on memory might be attributed to timing differences in the fluent and disfluent utterances: in the disfluent utterances, the *er* necessarily introduces more time between the context and the (predictable or unpredictable) target word. This might be particularly salient in the experimental situation, where many utterances end unpredictably. Among competing possibilities,

listeners could be sensitive to disfluent 'words' such as *er*, as suggested by Clark and Fox Tree (2002). Although the nature of the signal remains a question for future research (and some hints as to its resolution can be found in, e.g., Bailey & Ferreira's (2003) demonstration of the 'disfluency-like' effects of unnatural interruptions to speech), it is secondary to the primary motivation for the current study, which is to demonstrate that disfluent signals in speech affect listeners.

The effect of disfluency demonstrated in this paper is profound: differences in the processing of words in an utterance are visible immediately after the disfluency is encountered, and after a substantial delay (of up to 55 min after the first few utterances are heard) participants are more likely to recognise words which have been preceded by disfluency. Using a combined ERP and memory approach, we have established an effect of disfluency using a different type of predictability, and a different methodology, to those used by Arnold et al. (2004). Moreover, we have shown that the electrophysiological differences observed following hesitations are not merely epiphenomena, but reflect differences in immediate processing which have lasting effects. In other words, disfluency in speech has both short- and longer-term consequences for listeners.

Acknowledgements

We thank Jos Van Berkum for his helpful comments on an earlier version of this paper. Phil Collard, Alice Foucart, Graham MacKenzie, and Sinéad Rhodes each made useful suggestions.

This research was partially supported by a Staff Research Development Grant (M.C.), the Economic and Social Research Council (L.J.M.), and the Biotechnology and Biological Sciences Research Council (D.I.D.).

References

Arnold, J. E., Tanenhaus, M. K., Altmann, R. J., & Fagnano, M. (2004). The old and thee, uh, new: disfluency and reference resolution. *Psychological Science, 15*, 578–582.

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.

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

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

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.

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

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.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: brain potentials reflect semantic incongruity. *Science, 207*, 203–205.

Kutas, M., & Hillyard, S. A. (1984). Brain potentials during reading reflect word expectancy and semantic association. *Nature, 307*, 161–163.

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

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

McCarthy, G., & Wood, C. C. (1985). Scalp distributions of event-related potentials: an ambiguity associated with analysis of variance models. *Electroencephalography and Clinical Neurophysiology, 62*, 203–208.

Rugg, M. D., Mark, R. E., Gilchrist, J., & Roberts, R. C. (1997). ERP repetition effects in indirect and direct tasks: effects of age and interim lag. *Psychophysiology, 34*, 572–586.

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.

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*. Canada: Vancouver.

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.

Van Berkum, J. J. A., Brown, C. M., & Hagoort, P. (1999). Early referential context effects in sentence processing: Evidence from event-related brain potentials. *Journal of Memory and Language, 41*, 147–182.

Van Berkum, J. J. A., Zwitserlood, P., Hagoort, P., & Brown, C. M. (2003). When and how do listeners relate a sentence to the wider discourse? Evidence from the N400 effect. *Cognitive Brain Research, 17*, 701–718.

Van Petten, C. M., Coulson, S., Rubin, S., Plante, E., & Parks, M. (1999). Time course of word identification and semantic integration in spoken language. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25*, 394–417.