ON NOT REMEMBERING DISFLUENCIES

E. G. Bard and R. J. Lickley
Human Communication Research Centre and Department of Linguistics
University of Edinburgh, Edinburgh EH8 9LL, UK
Tel. +44 131 650 3951, E-mail: ellen@ling.ed.ac.uk

ABSTRACT

Disfluencies - repetitions and reformulations mid-sentence in normal spontaneous speech - are problematic for both psychological and computational models of speech understanding. Much effort is being applied to finding ways of adapting computational systems to detect and delete disfluencies. The input to such systems is usually an accurate transcription.

We present results of an experiment in which human listeners are asked to give verbatim transcriptions of disfluent and fluent utterances. These suggest that listeners are seldom able to identify all the words "deleted" in disfluencies. While all types suffer, identification rates for repetitions are even worse than for other types. We attribute the results to difficulties in recall or coding for recall items which can not be identified with certainty. This inability seems to make human speech recognition more robust than current computational models.

1. BACKGROUND

Human listeners are reasonably accurate in transcribing fluent speech but find it difficult to transcribe disfluencies [10]. In contrast, automatic speech recognition systems have considerable difficulty spotting and excising disfluencies despite the distinctive acoustic or structural features [3, 6, 11] of these very common phenomena. We report on the human abilities which make disfluencies evanescent.

The portion of a disfluent utterance which must be expunged to make fair copy is called the reparandum. Though words in reparanda are processed [4], they may not be correctly identified by normal listeners [8]. Part of the problem appears to be due to the disfluent interruption itself. People may depend on subsequent as well as prior context when they recognize words in running spontaneous speech [5, 2]. For words in the reparandum, the disfluent interruption truncates the subsequent context before the arrival of information which would normally allow the words to be identified. In a word-level gating experiment in which an utterance is presented starting with the first word and including an additional word on each trial, words in reparanda were so deficient in late-delivered recognition that they proved exceptionally unintelligible [8].

The current work examines the evidence that failures of memory as well as failures of perception are involved in the human ability to miss disfluencies. A large-scale verbatim transcription task was designed with two purposes. First, it checked for recognition failures in a more natural task than gating. Second, we test the hypothesis that REPETITION DEAFNESS will make recall even worse for disfluencies which contain repeated words than for those which do not. Repetition Deafness [9] and Blindness [7] are inabilities to distinguish in recall two very similar stimuli witnessed close together in time, particularly in presentations (e.g. rapid list intonation, time-compressed speech) which make perception and encoding difficult. We test two parts of this prediction: first, that the repetition itself creates the deficit, second, that coding and perceptual pressures conspire with repetition to suppress accurate reporting.

2. METHOD

2.1. Materials

Materials were spontaneous utterances from the HCRC Map Task Corpus [1], a set of task-oriented dialogs between pairs of undergraduate volunteers. Speech was digitally recorded in laboratory conditions with one stereo channel per speaker. Disfluencies were labelled and word-level segmentation performed via Entropic xlabel software with the aid of waveform and spectrographic representations.

Eighty simplex disfluencies, each containing a single contiguous reparandum, included 30 with no words from the reparandum repeated in the repair (hereafter, 'recasts'), and 50 with repetitions. The remaining 16 disfluencies were complex, containing either multiple attempts to repeat or to replace the reparandum or a series of different disfluencies. For 6 of these, the ultimate repair did not repeat any word in the preceding reparandum, while for the other 10, repetition was involved. For each disfluent utterance there was a fully fluent control utterance matching it for speaker and length in words.

As Table 1 illustrates, for each of the 96 disfluent utterances, four substrings were prepared. All began at

CHUNK	DISFLUENCY TYPE	
	REPETITION	RECAST
a	Right **there's a** {IP}	**There's about** {IP}
b	Right **there's a** {IP} *there's*	**There's about** {IP} *You've*
c	Right **there's a** {IP} *there's a*	**There's about** {IP} *You've got*
d	Right **there's a** {IP} *there's a* line about half way down	**There's about** {IP} *You've got* a yacht club right

Table 1. *Stimuli for two kinds of disfluency. Reparanda are in bold and repairs in italics. Interruption points ({IP}) were not indicated to listeners in any way*

the beginning of the utterance. The first (chunk a) ran up interruption point, the second (b) to the first word of the repair, the third (c) to the end of any repetition or, for non-repetition disfluencies, to the end of the next stressed word, and the fourth (d) to the end of the utterance. Control utterances were segmented at the corresponding positions. The substrings of each utterance were distributed by Latin square among four listener groups to give substring comparisons between subjects and fluency comparisons within subjects.

2.2. Procedure and subjects

Subjects were University of Edinburgh students with no known hearing loss. Nine were assigned to each listener group. Listeners were instructed to transcribe everything they heard into real words in the standard orthography and to be as accurate as possible even though some of the stimuli were difficult or odd. They were not told how many words any stimulus contained. Stimuli were presented three times in succession via high quality headphones. A transcription was required after each presentation.

3. RESULTS

We report analyses of first pass attempts at recall and transcription, the most natural listening condition.

As gating results would predict, listeners had great difficulty in reporting words from reparanda (Fig. 1). While control materials showed slight, insignificant improvement with longer stimuli, recall of words in reparanda was worse in the longest strings, where the completion of the utterance could often have allowed late recognition, than in the shortest strings, where only immediate recognition was possible (Fig. 2). Scoring whole reparanda and corresponding control words as right or wrong, the interaction between fluency and stimulus chunk length (a-d) was highly significant ($F_1(3, 105) = 122.10$, $p < .0001$; $F_2(3, 282) = 49.47$, $p < .0001$). All fluent outcomes were significantly better than any disfluent (Scheffés at $p < .01$). Recall for disfluent reparanda was significantly better in the stimuli (a) which stopped at the interruption point than in any of the longer substrings (at $p < .01$). The difference was not merely the effect of encountering a discontinuity at the point of

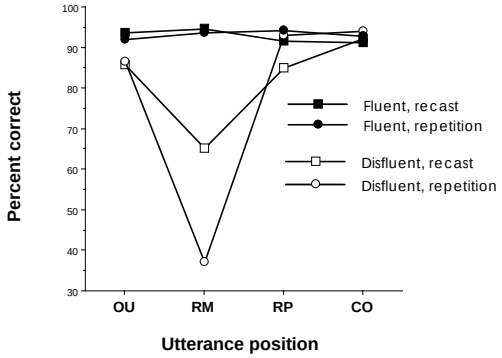

Figure 1. *Rate of correct report by fluency, part of disfluency (OU = Original Utterance - words before the Reparandum; RM = Reparandum; RP = Repair; CO = Continuation - words after the repair), and type of disfluency*

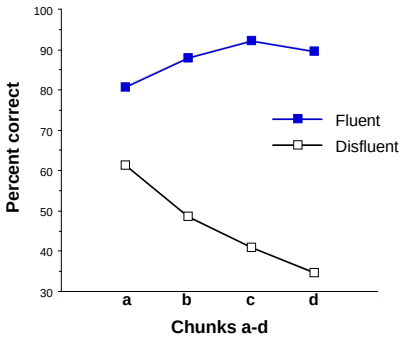

Figure 2. *Rate of correct report by substring length for words in reparanda of disfluent utterances and for corresponding words of fluent controls.*

interruption: chunk-d, the whole utterance, gave significantly worse recall than chunk-b, which included the first word of the repair.

Four kinds of evidence bear on the second hypothesis, that repetition deafness helps to expunge disfluencies. Repetition itself did not produce disastrous reductions in recall of the repeated word (at chunk-b for single-word

	REPETITION				RECAST			
CHUNK:	a	b	c	d	a	b	c	d
Multiple-R^2	.50	.47	.53	.52	.37	.43	.51	.52
Intercept	5.31	3.85	4.43	3.66	3.91	2.75	3.04	2.85
IPOS	**-.36**	**.45**	**.43**	-.21	-.23	**-.31**	**-.31**	**-.44**
CELEX	**.24**	.02	.03	-.10	**.23**	**.24**	**.30**	**.26**
BOUND	**.34**	**.29**	**.15**	**.31**	**.23**	**.24**	**.30**	**.26**
WDUR	-.04	.02	-.02	.07	**.41**	**.35**	**.35**	**.27**
IPWW	-.24	-.05	**-.34**	-.23	**-.31**	**-.30**	**-.39**	**-.33**

Table 2: *Contributions to multiple regression equations for words in reparanda at each stimulus length (a-d), significant predictors only. Beta values in boldface have $p < .05$.*

reparanda, chunk-c for others). Instead, stress on memory or processing load seemed to promote special deficits for repetition disfluencies.

First, there is a repetition deficit. For the most vulnerable words, those just preceding the interruption point, report rate falls more sharply in repetition disfluencies than in others. In an analysis of recall loss, i.e., how much recall rates changed from the chunk-a (ending just after the words assessed) to the later chunks, the down-turn for repetitions was particularly marked: (fluency x chunk x disfluency type: F_2 (1, 367) = 4.50, $p < 0.035$). Again, the fall-off in recall continued beyond the point where the repetition occurred (chunk-b/c) and so may be due to the memory load created by the additional words in chunk-d. Fluent controls showed no comparable trends.

Second, extensive exploration of the results by multiple regression analyses showed that recall for repetition and recast disfluencies were subject to somewhat different influences. All words from all disfluent stimuli were coded for dictionary characteristics of the words (CELEX = database raw frequency, FUNCONT = functor or contentive word class membership), for their characteristics as uttered tokens (BOUND = strength of following phonological boundary, from sentence boundaries at 3 down to functor-contentive boundaries at 0; WDUR = msec length of word; PDUR = msec length of following pause; STRESS from 2 for pitch accent to 0 for no stress), and for characteristics of their location in a disfluent utterance (IPOS = the length of chunk-a; IPWW = distance of word from disfluent interruption point, RMWW = number of words in the reparandum, UTWW = number of words in utterance). All words from fluent stimuli were coded for the same variables with characteristics of the matched disfluent partner used for certain position variables (IPWW, IPOS, RMWW).

In a set-hierarchical multiple regression, equations including characteristics of the disfluent utterance always accounted for significantly more of the variance in recall rate than equations lacking these variables . Table 2 displays the significant contributors in a typical set of equations for reparandum words and their controls.

All words showed effects of the structure of their disfluency: proximity to the interruption and longer sequences of words before the interruption point made for worse recall. All showed the influence of structure which we know affects prompt recognition in fluent speech: words preceding more important prosodic and syntactic boundaries were reported better. After chunk-a, however, only recast words depended on variables which would have made them more intelligible out of context (length and frequency), though repetition and recast words have similar means and ranges for these variables. Recall of words in repetition disfluencies appears to be largely dependent on the surrounding structures, rather than on the words themselves, while the words which need to be expunged from recasts persist if they are easy to recognize.

Third, we can see a direct effect of the complexity of the utterance if we examine results for the final word in the reparandum. We compare recall for simplex disfluencies, where the utterance is fully fluent up to the interruption point, with complex disfluencies, where multiple interruptions disrupt the string (Fig. 3). In simplex single-word reparanda, repetition and other disfluencies behaved alike ($F_2 < 1$). As in the earlier analyses, fluent control words were somewhat easier to report when more context was presented, while reparandum-final words were reported less accurately in longer strings (fluency x chunk: $F_2(2, 68) = 14.06$, $p < .0001$). For the complex disfluencies, there is both a detrimental effect of longer stimuli ($F_2(2, 28) = 10.88$, $p < .0003$) and an additional deficit for repeated words (disfluency type x fluency x chunk: $F_2(2,28) = 3.81$, $p < .035$). In other words, repetition disfluencies are significantly more forgettable than others when they occur in utterances which are already difficult to process because of multiple false starts.

The final evidence for repetition deafness as a function of processing pressure is the difference between the current results and those reported on the same materials under word-level gating [8]. In the latter technique, word boundaries are indicated since each word forms the end of some stimulus, and subjects hear very little new material on each trial. Words in repetition disfluencies

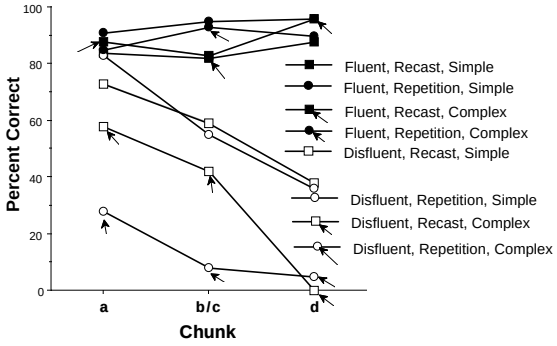

Figure 3. *Rate of correct report for final word of reparandum and for corresponding word of fluent controls by fluency, complexity of reparandum, and type of disfluency.*

were recognized somewhat better as subsequent context accrued. In the transcription technique, where word boundaries were not marked, and recognition of many words was required on a single trial, the same disfluencies show significantly more tendency to suffer from additional context (fluency x task: $F_2(1,388) = 10.16$, $p < .01$).

4. CONCLUSIONS

The results of the transcription study indicate that normal listeners are seldom able to recognize words in reparanda, particularly words close to the interruption point, when they listen to whole utterances or long substrings. The effect is more severe when stimuli include more words after the interruption point. It is more marked still if the disfluency involves a repeated word.

We attribute these results to the difficulties of recalling or coding for recall those items which are difficult to identify with certainty. Words in both kinds of disfluencies were harder to report if they occurred in complex reparanda: all disfluent material was vulnerable in situations where context made it difficult to identify. Repeated material was especially susceptible to context effects from the length and structure of the carrier utterance and from the integrity of the reparandum itself. It is attractive to associate these results with repetition deafness phenomena reported for other speech. These phenomena occur only when there is unusual pressure on the perceptual processes, but there is considerable dispute as to whether they are caused by initial failure to discriminate one stimulus from another or by encoding and retrieval processes. Certainly the evanescence of disfluencies suggests that initial perceptual difficulties create material which cannot be adequately coded in human memory on a single presentation. Human inabilities are conveniently adaptive in this instance.

5. REFERENCES

[1] A.H. Anderson, et al. “The HCRC Map Task Corpus”. *Lang&Speech*, **34**:351-366, 1991.

[2] E.G. Bard, R.C. Shillcock & G.T.M. Altmann. “The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context”. *Perc&Psychophys*, **44**(5):395-408, 1988.

[3] J. Bear, J. Dowding & E.E. Shriberg. “Integrating multiple knowledge sources for detection and correction of repairs in human-computer dialog”. *Proc. ACL*, 1992.

[4] J.E. Fox Tree. “The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech”. *JML*, **34**:709-738, 1995.

[5] F. Grosjean. “The recognition of words after their acoustic offset: Evidence and implications”. *Perc&Psychophys*, **38**(4):299-310, 1985.

[6] P.A. Heeman, K. Loken-Kim & J. Allen. “Combining the detection and correction of speech repairs”. *Proc ICSLP 96*: 362-365, 1996.

[7] N.G. Kanwisher. “Repetition blindness: type recognition without token individuation”. *Cognition*, **27**:117-143, 1987.

[8] R.J. Lickley & E.G. Bard. “On not recognizing disfluencies in dialogue”. In *Proc ICSLP 96*: 1876-1879, 1996.

[9] M.D. Miller.& D.G. Mackay. “Relations between language and memory - the case of repetition deafness”. *PsychSci* **7**(6):347-351, 1996.

[10] J.G. Martin & W. Strange. “The perception of hesitation in spontaneous speech”. *Perc&Psychophys*, **3**(6):427-438, 1968.

[11] C. Nakatani & J. Hirschberg. “A corpus-based study of repair cues in spontaneous speech”. *JASA*, **95**:1603-1616, 1994.