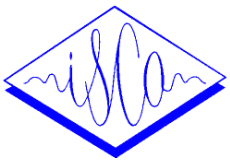


ITRW on Disfluency in
Spontaneous Speech (DiSS'01)
Edinburgh, Scotland, UK
August 29-31, 2001

Is disfluency just difficulty?

Ellen G. Bard, Robin J. Lickley, Matthew P. Aylett

Department of Theoretical and Applied Linguistics and Human Communication Research Centre
University of Edinburgh, Scotland
ellen@ling.ed.ac.uk

Abstract

The question addressed by this paper is whether disfluency resembles Inter-Move Interval, a measure of reaction time in conversation, in displaying effects of the overall difficulty of conducting a coherent conversation. Five sources of difficulty are considered as potential causes of disfluency: planning and producing an utterance, comprehending the prior utterance, performing a communicative task, order effects, and interpersonal factors. A multiple regression analysis on simple disfluencies in the HCRC Map Task Corpus shows that planning and production make the major independent contribution to predicting the rate of disfluencies, with interpersonal variables and position in dialogue also contributing significantly. Notably, comprehension variables did not affect either the total rate of disfluency or the rate of individual kinds of disfluencies.

1. Introduction

Many disfluencies are edited errors in speech production. They mark those occasions when speakers have not framed an utterance which satisfies their goals before they begin to speak. Disfluencies are thought to occur when speakers fail to monitor and edit successfully during earlier phases of production [1, 2]. We do not yet know exactly what prevents correct initial formulation or internal self-correction in natural circumstances, but there are many possible culprits among the tasks competing for the speaker's attention. To produce any spontaneous utterance, a speaker must plan, assemble, and articulate a string of words. In dialogue, interlocutors must also comprehend one another's contributions and provide appropriate replies promptly enough to make it plain that they wish to take the floor. In task-oriented dialogue, they must use the interaction to achieve a non-conversational goal. As with any other task, initial attempts at any of these activities in a given setting may prove difficult.

If disfluency is induced by such difficulties, then it should behave like Inter-Move Interval (IMI). Defined as the time, positive or negative, between the offset of one speaker's utterance and the onset of the interlocutor's reply, IMI is a measure of reaction time in dialogue [3]. IMI is longer early in a dialogue, when the interlocutor's utterance is difficult to comprehend, when the interlocutors are having difficulty with the task, when a long utterance follows, and when that utterance begins a larger unit of dialogue. At the same time, IMI is shorter in what might be the more delicate social situation: conversations between persons of different sexes who have just met. Thus, time to begin speaking is sensitive to interpersonal factors as well as to cognitive pressures of various kinds.

There is already evidence that planning and production burdens affect fluency. Disfluencies tend to occur early in utterances, when planning of later stages is incomplete. Disfluencies are more common in longer utterances [4-6], in more complex constituents [4], and when response choices are complex [6].

There is good reason to predict what affects delay to begin speaking will also affect fluency of speech. First, we predict effects o*f **the prior utterance***. Pressure to hold or take the floor may induce imperfections in planning [5,7]. Since modal IMI, at around 150 msec, with 14% negative, is too low to allow planning to follow listening entirely [3], speakers must begin planning their next utterance while listening to their interlocutor's prior utterance. If normal comprehension processes are also used for self-monitoring [8], then any competition should obstruct production. The effect might be more disfluency at shorter IMIs or more disfluency with longer or more complex prior utterances. Also, the fluency of the prior utterance may be important: cross-speaker syntactic priming [9] or mere difficulty in perceiving disfluent utterances [10] could yield disfluent adjacency pairs. Second, we might predict effects of ***task difficulty*** in general, because competition for attention is likely to affect any process serving an activity as complex as conversation. Third, if task-oriented dialogue comprises as a series of similar problems which have to be solved by communicating, then there may be ***order effects*** within a single such conversation or across a series. Certainly, dialogues and expressions used in them get shorter with time [11]. By building expertise and mutual knowledge, interlocutors effectively narrow the choices they have to make, and these basic conditions should enhance fluency. Finally, it seems likely that ***interpersonal factors***, like the number of sensory channels or the familiarity of the interlocutors should affect delicate processes of planning and feedback [12]. We know that these variables affect the structure of dialogues where sensory channels or communicative links are limited (e.g., [13]).

The difficulty with testing so many predictions, of course, is that the predictors may intercorrelate. The longer utterances earlier in a conversation, for example, may induce disfluencies because they are long, because they are early, or both. To determine what independent contributions are made by each kind of predictor, we use a method similar to the one which found predictors of IMI [3]: we run a multiple regression analysis on all appropriate items from the same coded corpus of task-oriented dialogues, deriving our reported results from the most fully coded subset of the materials.

2. Method

2.1. Corpus

Materials came from the HCRC Map Task Corpus [14] (hereafter MTC), 128 unscripted dialogues in which 32 pairs of Glasgow University undergraduates communicated routes

defined by labeled cartoon landmarks on schematic maps of imaginary locations. Instruction Giver's (hereafter 'IG') and Follower's (IF) maps for any dialogue matched only in alternate landmarks. Participants knew that their maps might differ but not where or how. Players could not see each other's maps. Familiarity of participants (within subjects) and ability to see the interlocutor's face (between subjects) were counterbalanced. Each participant served as IG for the same route to two different IFs and as IF for two different routes. Channel per speaker digital stereo recordings were orthographically transcribed and digitally word-segmented. Like the coding systems described below, the segmentations form part of an XML corpus database.

2.2. Unit of analysis

The word-segmented corpus is framed as a series of Conversational Game Moves [7], turns or parts of turns whose purpose in moving the dialogue forward can be determined by their form and context. Moves are stages of Conversational Games, which are themselves usually stages in completing Transactions, sections of the task which the dialogue serves [15]. Here we used only those Moves which involve a change of speaker and which are likely to be a reply to the previous speaker's Move: we excluded those which began too early to respond to the prior Move (onset preceding offset of the prior speaker's Move by > 1 sec, or onset < 350 msec after prior Move onset) and those which were actually resumptions of an earlier Move by the same speaker (onset < 300 msec after the end of a previous Move by the speaker).

2.3. Disfluency coding

The dependent variables were ***numbers of disfluencies*** of various kinds ***per Move***. Disfluency annotation [16] was performed on the whole corpus using Xwaves/Entropic xlabel. Annotators examined the speech waveform closely and made use of spectrograms where necessary. For each disfluency, individual words were labeled by part and type of disfluency. Disfluency parts [17] are original utterance, reparandum, interruption/filler, repair and continuation.

The current paper omits silent and filled pauses, combination and complex disfluencies, and reports on only simple disfluencies of 4 types. In ***repetitions*** the speaker repeats a string verbatim, with no additions or deletions: e.g. *[we're going] we're going left of the camera shop.* In ***insertions*** the speaker repeats a string and inserts a word or words within the repeated string: e.g. *[go left] go just left of the camera sho.* In ***substitutions*** a word or string is replaced by another with no major syntactic alteration: e.g. *go [left] right of the camera shop*. In ***deletions***, the speaker interrupts an utterance and either restarts without repeating or directly substituting or simply surrenders the floor to the other speaker: e.g. *[you're away f-] right see the wee bit that's jutting out?*

2.4. Predictor variables

2.4.1. Current Move

Three sets of predictors reflect hypotheses about how production tasks encourage disfluency. First, the planning functions are represented by the ***speaker's role*** (because Instruction Givers bear more of the burden of structuring the dialogue), and by the conversational ***boundary*** preceding this current Move. If planning a section of the task or dialogue affects fluency as it affects IMI, then disfluency rates should follow the size of the planned unit, with Transaction-initial Moves (see 2.2) most disfluent (2), Game-initial Moves (1) somewhat less disfluent, and Game-internal Moves (0) least disfluent . Second, the burden of constructing ***referring expressions*** [4] is measured via separate counts for the combinations of New/Given and Shared (on both players' maps) / Unshared (on only one). Finally, as a more general indicator of complexity, ***length in words*** is measured, but omitting any words in the reparanda of disfluent Moves.

2.4.2. Prior Move

Three sets of predictors represent aspects of the prior speaker's utterance which may make the current speaker disfluent. First, difficulty in comprehending a complex set of references to map locations may interfere with the process of production. Hence, numbers of ***referring expressions*** in the prior Move are classed as for current Moves (2.4.1). To test for priming by disfluent structures, prior Move ***disfluency*** is tallied for each kind of disfluency (see 2.3). Finally, ***length in words*** is included.

2.4.3. Difficulty metrics

To reflect difficulty in pursuing the task itself, we use 3 measures. ***Deviation score*** is the mismatch in cm^2 between the model route on IG's map and the route ultimately drawn on the IF's. Major miscommunications yield large deviation scores. ***Drawing*** shows whether the prior Move was followed by an attempt to draw part of the route. Finally, ***Inter-Move Interval*** itself is an indicator of various kinds of cognitive load [3].

2.4.4. Order

To capture effects of practice and of increasing discourse context, 2 order codes were used. ***Conversation*** records which of the 8 dialogues produced by a quad (pair of speaker pairs) is in progress. Conversations 4-8 are second trials with a map on the part of the IG. ***Position*** is the ordinal position of the Move in the dialogue (μ = 136.69, *s.d.* = 110.97).

2.4.5. Interpersonal

These are aspects of the corpus design which affect the social distance between IG and IF. ***Eye-contact*** refers to the presence (0) or absence (1) of a flimsy barrier blocking the line of sight between interlocutors. ***Familiarity*** records whether the pair have just met (0) or are friends (1).

3. Results

Detailed results are reported for the 6882 'response' Moves (see 2.2) of the dialogues coded for the presence of drawing between Moves. The results were essentially the same for the whole corpus, that is all 14389 'response' Moves not containing complex disfluencies (see Figures 1-7).

3.1. Significant contributions

Multiple regression equations using all predictors were prepared with total number of disfluencies per move as dependent variable, and then with number of each individual

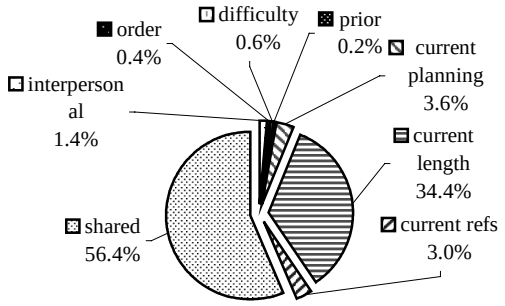


Figure 1.Contributions of groups of predictor variables to the explained variance in total rate of disfluencies

subcategory. Table 1 displays only the significant β-values (standardized regression coefficients) for total disfluency count, since the patterns of results are the same throughout. All the regression equations accounted for significant proportions of the variance in disfluency rates ($p < .0001$), with explained variance in the overall measure (Multiple R^2) nearly 14% (deletions 2.5%, insertions 5.3%, repetitions 8.1%, substitutions 4.3%).

The principal question addressed was whether disfluency behaves like IMI in sensitivity to cognitive load from many sources. As Table 1 shows, it does not. Significant individual predictors are restricted to characteristics of the current utterance and to interpersonal and order effects. Difficulty and prior Move variables do not make significant individual contributions to accounting for the rates of disfluencies. Figure 1 displays the proportion of accounted-for-variance attributable to each group of variables. As expected, a large proportion of the explained variance (> 56%) is shared among the intercorrelating predictors. Of the five groups, the current Move predictors clearly predominate (41%), with length in words alone accounting for over 30% of the variance in total disfluency rate, more than all other groups combined.

3.2. Individual effects

The remaining figures display effects of individual predictor variables on overall disfluency rate. Raw means are used for simplicity of interpretation, but significant β values indicate that trends would be robust even adjusted for effects of other predictors.

Figures 2 and 3 display predicted effects of planning on disfluency. Figure 2 shows that IGs, who usually take responsibility for directing the dialogue, are more disfluent (.218) than IFs (.093) (β = .07, p < .01). Figure 3 shows that there are more disfluencies in Moves which initiate larger constituents of a dialogue, with rates of disfluency rising from Game-internal Moves (.116) to Game-initial (.219) and again to Transaction-initial (.294) (β = .02, p < .05).

Figure 4 and 5 show the predicted effects of current-Move length and referential complexity. As in other corpora [4, 6], longer Moves attract more disfluencies (rising continuously from .063 for single-word Moves to .709 for > 17 words, β = .28, $p < .01$) (Figure 5). Moves containing more referring expressions (.104 for 0, .248 for 1, .492 for 2 to 5) also attract more disfluencies (β = .28, $p < .01$).

Figure 6 shows an order effect but clearly not a simple practice effect: Moves later in the dialogue exhibit higher rates of disfluency (.147, .135, .169, .167 for successive quartiles, β = .03, $p < .05$).

Finally, Figure 7 associates disfluency with interpersonal difficulty: speech to an unfamiliar partner is more disfluent (.174 v .140, β = -.03, $p < .05$).

Table 1. Significant β-values in multiple regression equations predicting occurrence of disfluencies in 6882 Conversational Game Moves (coded for presence of drawing). df *= 23, 6858, p < .0001. (Key: * :* p < *.05, ˚:* p < *.01)*

Type		Variable	All
Interpersonal		Familiarity	- .04˚
		Eyecontact	
Order		Conversation	
		Position	.03*
Difficulty		I.M.I. deviation score	
		Drawing	
Prior move	Reference	new shared	
		given shared	
		new unshared	
		given unshared	
	Disfluency	deletions	
		repetitions	
		substitutions	
		insertions	
	Length	length (words)	
Current move	planning	Role	.07˚
		boundary	.02*
	reference	new shared	
		given shared	.06˚
		new unshared	.02*
		given unshared	.03*
	length	length (words)	.28˚
		Multiple R^2	.138
		F	47.64
		n	1065

Figure 2 . Effects of role on disfluency per move

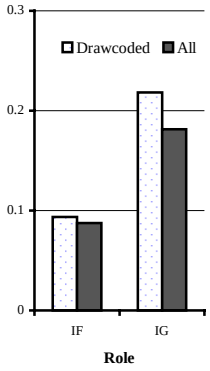


Figure 3. Effects of dialogue unit on disfluency per move

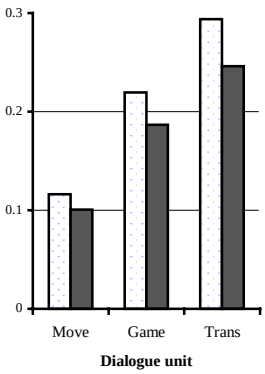


Figure 4. Effect of Move-length on disfluency per move

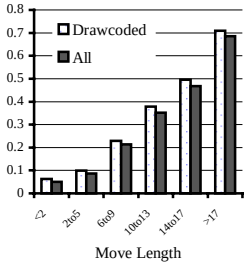


Figure 5. Effect of referring expressions on disfluency per move

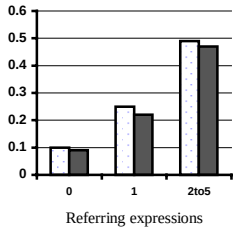


Figure 6. Effect of position in dialogue

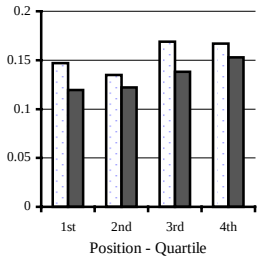
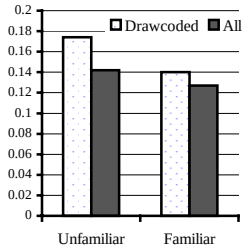


Figure 7. Effect of familiarity



4. Conclusions

The results of the multiple regression analyses of the correlates of disfluency show quite a different pattern from the one observed for IMI. Rather than behaving like a general measure of difficulty affected by interpersonal, practice, comprehension and production processes, disfluency seems to be linked to processes of production, with the greater part of the uniquely explained variance attributable to characteristics of the disfluent Move: length, referential complexity and likely role in larger scale planning of the dialogue.

The smaller contributions of unfamiliarity and position in dialogue could also be construed as difficulty effects: the theory of common ground suggests that framing a satisfactory utterance may be more difficult if the addressee is a stranger rather than a friend. The effect of position in dialogue, here measured by Move number, may be unduly influenced by dialogues which are unusually long because communication is proving difficult..

Yet it is plain that disfluency has a particular area of insensitivity: Even though human language production and comprehension are thought to share components, disfluent output is not associated with any of the current measures difficult or disfluent input. This fact suggests a separation rather than a sharing of processes.

5. Acknowledgments

This work was supported by the ESRC main grant to HCRC and by EPSRC Project Grant GR/L50280/01. Dr Aylett is now with Rhetorical Systems, Edinburgh.

6. References

[1] Blackmer, E., and Mitton, J., "Theories of monitoring and the timing of repairs in spontaneous speech", *Cognition*, 39:173-194, 1991

[2] Postma, A. and Kolk, H. "The effects of noise masking and required accuracy on speech errors, disfluencies, and self-repairs", *J. Speech Hearing Res*., 35:537-544, 1992.

[3] Bard, E. G., Aylett, M., and Bull, M. "More than a Stately Dance: Dialogue as a Reaction Time Experiment", *Proc. Soc Text and Disc.*, 2000.

[4] Clark, H. H., and Wasow, T., "Repeating words in spontaneous speech", *Cognitive Psychology*, 37: 201-242, 1998.

[5] Maclay, H., and Osgood, "Hesitation phenomena in spontaneous English speech", *Word*, 15:19-44, 1959.

[6] Oviatt, S., "Predicting disfluencies during human-computer interaction", *Comput. Speech Lang.*, 9:19-35, 1995.

[7] Carletta, J. , Isard, A. , Isard, S. , Kowtko, J. , Doherty-Sneddon, G. , Anderson, A., "The reliability of a dialogue structure coding scheme", *Computat. Ling.*, 23:13-31, 1997.

[8] Levelt, W.J.M., Roelofs, A., and Meyer, A. S., "A theory of lexical access in speech production", *Behav. Brain Sci*., 22:1-45.

[9] Branigan, H., Pickering, M., and Cleland, A., "Syntactic coordination in dialogue", *Cognition*,, 75:813-825, 2000.

[10] Bard, E. G., and Lickley, R. J. "Graceful failure in the recognition of running speech", *Proc 20th Ann. Meeting of the Cog. Sci. Soc.*, 108-113.

[11] Brennan, S., and Clark, H., "Conceptual pacts and lexical choice in conversation", *JEP:LMC*, 22:1482-1493, 1996.

[12] Branigan, H., Lickley, R., McKelvie, D. Non-linguistic influences on rates of disfluency in spontaneous speech. Proc. 14th ICPhS, 1999.

[13] Doherty-Sneddon, G., Anderson, A. H., O'Malley, C., Langton, S., Garrod, S., and Bruce, V., "Face-to-face and video-mediated communication: A comparison of dialogue structure and task performance" *J Exp. Psych.: Applied*, 3:105–125, 1997.

[14] Anderson, A., Bader, M., Bard, E.G., Boyle, E., Doherty, G., et al., "The H.C.R.C. Map Task Corpus", *Lang. and Speech*, 34:351-366, 1991.

[15] .Isard, A., and Carletta, J., *Transaction and Action Coding in the Map Task Corpus*. Research paper HCRC/RP-65. HCRC, U. of Edinburgh., 1995.

[16] Lickley, R.J., HCRC *Disfluency Coding Manual*. Technical Report HCRC/TR-100, HCRC, U. of Edinburgh., 1998.

[17] Levelt, W.J.M., "Monitoring and self-repair in speech", Cognition, 14:14-104, 1983.