ISCA Archive
http://www.isca-speech.org/archive

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

Dialogue moves and disfluency rates

Robin J. Lickley

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

Abstract

Many factors conspire to cause speakers to produce hesitations and self-repairs in dialogue. It has been noted that disfluency rates vary between corpora, with different overall dialogue tasks and with different modalities (e.g. human-computer vs. human-human) and between speakers, where they play different roles within a given dialogue.

In this paper, we attempt to account for some of these results by examining the interaction between rates of different types of disfluency and types of utterance (dialogue moves) within one corpus of human-human task oriented dialogues.

We find both that overall disfluency rate varies by dialogue move type, with moves which require more planning producing more disfluency, and that the distribution of disfluency types varies between move types, most notably with complex and negative responses to questions producing more filled pauses than positive replies and other moves.

This work helps us to understand how dialogue structure can account for differences in disfluency rates between and within speech corpora and has implications for research in speech production and perception, discourse studies, dialogue management and automatic speech recognition.

1. Introduction

As if understanding normal spontaneous speech wasn't difficult enough, we have to make it tougher, by throwing in disfluencies with alarming frequency. So, why be disfluent? Two major functions of disfluency are **hesitation** and **self-repair**.

Speakers may hesitate before or during an utterance for any of a number of reasons. Hesitation may take the form of pausing, either via silence, filled pause (e.g. *em uh, um*) or word prolongation or some combination of these, or it may occur as repetition of utterance onset.

- Planning what to say next takes time, whether it be a matter of formulating the whole utterance, or selecting the most appropriate lexical item, searching memory for a word or to answer a question [1], or physically searching for information needed to complete an utterance (e.g. *his phone number is ...*} – speaker pauses while looking up a number).
- If an error has been detected, whether overtly or covertly (i.e. prior to articulation: e.g. [2,3]), it may take time to replan the utterance and produce the repair.
- In dialogue, a speaker may hesitate to ensure that their interlocutor is paying attention [4] and not speaking at the same time or attending to some other task or to ensure that they can be heard when competing with extraneous noise.
- Also within dialogue, the realisation that there is a mismatch between interlocutors' representations of the current dialogue state [5] may necessitate hesitation for replanning purposes.

Speech production is a dynamic act which involves more or less constant planning and self-monitoring on many levels. Spontaneity results in errors, and in the vast majority of cases, speakers detect and correct their own errors, producing self-repairs. Levelt [2] distinguishes two major classes of self-repair:

- in *appropriateness repairs*, the speaker realises that the message needs to be made more specific in some way, for the meaning to be conveyed correctly (e.g. *go **past** - just a short distance past the tree*).
- in *error repairs*, the speaker detects a mistake at some level of production and this needs to be corrected (e.g. *go **path** - go past the tree*).

As with hesitation, speakers may be forced into performing repairs in order to adjust for their view of the hearer's knowledge state (e.g. *turn left at the po- do you know where the post office is?*).

While most disfluencies fall into the categories of hesitation and repair, some cases may simply be habitual. The "RP stutter", for example, multiple repetitions of utterance-initial words is fairly common in certain groups of Southern British English.

Previous work has given us insights into how disfluency rates vary within and between speech corpora.

Within corpora, role, sex, eye-contact and familiarity have been found to have some effects on disfluency rates. We find higher disfluency rates for speakers whose role it is to give instructions compared to their interlocutors [7,8,9]. There is some evidence that male speakers may be more disfluent than females [7,8,9,10,11]. Work on our corpus suggests that there are some effects of eye-contact on disfluency rates, with higher rates of repetitions when people are unable to see each other [7], and effects of familiarity, with speakers who do not know each other producing more disfluency [9].

Uncertainty in answering questions has been found to result in greater use of filled pauses within general knowledge tests [1,12], though no comparison was possible with other types of utterance.

Syntactic complexity has effects on repetition rates, with complex structures more likely to be prefixed by repetition disfluencies than simple structures [13].

Discourse structure has also been found to interact with disfluency rates. In monologues, disfluencies cluster around ideational segment onsets (e.g. [14]). Within dialogues, we have a similar finding, that utterances which commence larger dialogue units contain more disfluencies [9]. Longer utterances have been found to be associated with higher rates of disfluency [15], but not for all corpora [11].

Other cross-corporal differences include higher disfluency rates for human-human than for human-computer dialogues, higher rates in telephone conversations than face-to-face and lower rates when dialogues are more constrained [11,15].

Various of the works cited above have taken into account different types of disfluency. Some have looked at dialogue structure. This work is the first that attempts an analysis of a large corpus of spontaneous dialogues, taking into account variations in dialogue move types and disfluency types. We ask what kinds of dialogue moves are most likely to contain disfluencies, whether any differences remain when we take into account length of utterance and whether different move types attract different disfluency types.

2. Corpus and methods

The materials for this study come from the HCRC Map Task Corpus (hereafter, MTC) [16], a collection of 128 dialogues between 64 Scottish undergraduates, comprising about 15 hours of spontaneous dialogue and around 150,000 words. In each dialogue, speakers take the role of instruction Giver or instruction Follower, the Giver describing a route through a map to the Follower who has the same basic map but a slightly different set of landmarks to negotiate. The entire corpus has been transcribed orthographically, with individual words and intervals between words time-aligned with the speech signal and annotated at many levels. Importantly, the annotation schemes are in a format (XML) which allows us to make enquiries about interactions between the various levels [17].

2.1. Move annotation

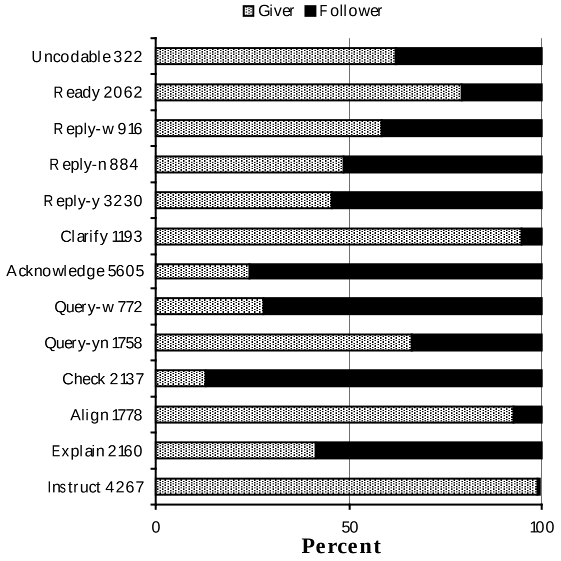
The MTC is fully annotated for three levels of dialogue structure: **transactions** are subdialogues which form major units in the overall task (typically, in the MTC, a transaction involves the completion of one segment of the route on the follower's map); **games** are components of transactions and comprise sets of dialogue exchanges which achieve a specific sub-goal (e.g. successful completion of an instruction); **moves** are the utterances which make up games (e.g. an instruction game, may consist of the following set of moves: instruct, query, reply, explain, align). Details of the dialogue annotation are given in [18,19].

We focus on *moves* in this paper. Moves fall into the three subcategories of initiation, response and preparation. Initiation moves are commands (INSTRUCT), statements (EXPLAIN) or questions (ALIGN (e.g. *have you got that?*), CHECK, QUERY). Response moves may show that a command has been successful (ACKNOWLEDGE) or provide an answer (REPLY) or an amplified answer (CLARIFY). . "Uncodable" moves are shown in Figure 1, but will be omitted from further analysis in this paper, as will the single type of preparation move (READY).

2.2. Disfluency annotation

The whole MTC is annotated for disfluencies, following the Map Task Disfluency Coding Manual [20]. Disfluencies are labelled for type, (delete, repetition, insertion, substitution, combinations of these with a single interruption point), for length of reparandum (in words), for complexity (embeddings), and for speaker-overlap. Individual words within disfluent reparanda and repairs (c.f. [2]) are labelled for their role within the disfluency and silent and filled pauses and editing expressions associated with the main disfluency types are marked. Annotation was performed against the speech wave-forms and word-level transcriptions, using Xwaves and Xlabel from Entropic and the resulting files were converted into XML.

Figure 1. Move type (and number in whole corpus) by speaker role



2.3. Data extraction

Data for Moves and Disfluency were extracted from the XML version of the MTC via unix-based xml-query tools [21] as well as standard UNIX tools.

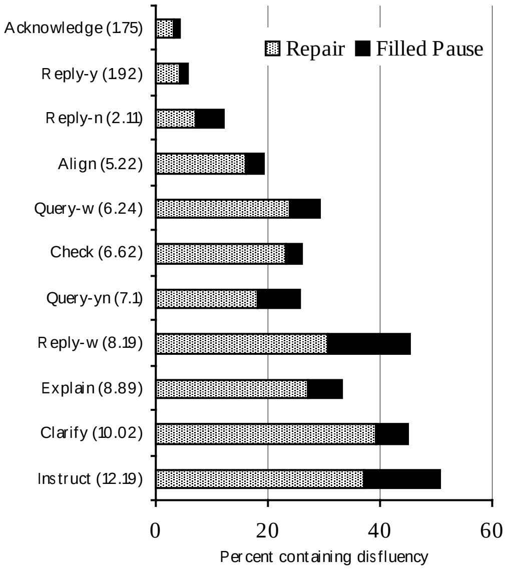
3. Results

Figure 1 shows the count and distribution by speaker role for move types in the whole MTC. We exclude uncodable and "ready" move types from further analysis and focus on the results from the remaining 24,697 moves, containing 2,249 deletions, 1675 repetitions, 555 insertions, 712 substitutions, 293 combination types (all of which we will refer to as "repair-type disfluencies" henceforth) and 1491 filled pauses.

The first observation is that all disfluency types occur in all move types. But, as Figure 2 suggests, move types vary considerably (1) as to whether they are likely to contain any disfluency at all and (2) as to their ratio of repairs to filled pauses. 43.4% of the longest moves (**Instruct**) are disfluent, while only 4.2% of the shortest (**Acknowledge**) are. A larger proportion of disfluent **reply-n**, **reply-w** and **instruct** moves have filled pauses than other moves. Of course, these preliminary observations need to be enhanced by taking into account disfluency rate per words and utterance length, and we do so in the analyses that follow.

We take as our dependent variables, disfluency rate per 100 words and report rates for each type of disfluency introduced above, as well as total rate for repair-types, filled-pause rate and overall disfluency rate. In this study, we include in the word count words in reparanda, but not filled pauses. The results reported here are for General Linear Model multivariate analyses with the Move type (the 11 types shown in Figure 2) as independent variable. Post Hoc *Tukey* tests are used to show differences in disfluency rates by homogeneous subgroups of move types – types within each group do not differ significantly from other group members.

Figure 2. Percent of moves containing repairs and filled pauses, ordered by mean length in words (N).



For the full set of data, rates of all disfluency types vary significantly by move type (p < .001). Table 1 shows homogeneous subsets of moves for overall disfluency rates. For this data, replies to wh-questions (**reply-w**) stand out as the most disfluent move type, followed by **clarify** (amplified replies), **instruct**ion and wh-question (**query-w**) moves. The least disfluent - **acknowledge**, positive replies (**reply-y**) and **align** moves – are amongst those with the shortest mean length in words.

Table 1. Homogeneous subsets of move types for total disfluency rates – cells show mean disfluency rates. (α=.05)

Move Type	Subset					
	1	2	3	4	5	6
Acknowldg	1.74					
Reply–Y	1.93					
Align	2.52	2.52				
Query-YN		3.77	3.77			
Reply-N			4.51	4.51		
Explain			4.66	4.66		
Check			4.81	4.81		
Query-W				5.49	5.49	
Instruct					6.29	
Clarify					6.59	
Reply-W						8.22

That some of the shortest moves are in the subset with the lowest disfluency rate and some of the longest moves are in the higher subsets suggests that there may still be an effect of move length on disfluency rates (although, in fact, there is not a simple correlation between move-length and disfluency rate in this data). So, for more detailed analysis, we take that subset of the data that includes only moves of between 4 and 6 words in length, comprising 4,973 moves with a minimum of 107 moves and a maximum of 819 moves of any one type.

For the length-controlled subset of the data, the overall picture is similar to the full set. All but the numerically smallest type of disfluency (combination) differs significantly for rate by move-type (p<.05). For total disfluency rates, the moves fall into 3 homogeneous subsets (Table 2). Three facts are notable in the move-type subgroups. (1) 3 of the 4 most disfluent moves represent either complex or negative replies to questions. (2) 2 of the move disfluent subgroup are predominantly Giver moves (see Figure 1), while the others are role-neutral. (3) With shorter moves removed from the analysis, move-types which for the whole data set are predominantly 1 or 2 words long (**Acknowledge** and **reply-y**) no longer stand out as having vastly smaller disfluency rates.

Table 2. Homogeneous subsets of move types for total disfluency rates for moves of 4-6 words in length – cells show mean disfluency rates. (α=.05)

Move Type	Subset		
	1	2	3
Query-YN	2.12		
Align	2.43		
Reply–Y	2.91		
Explain	3.08		
Acknowldg	3.41	3.41	
Check	3.63	3.63	
Query-W	3.66	3.66	
Clarify		5.51	5.51
Reply-N		5.59	5.59
Instruct			6.33
Reply-W			6.47

Figure 3 . Repair and filled pause rates for moves of 4-6 words, ordered by overall disfluency rate.

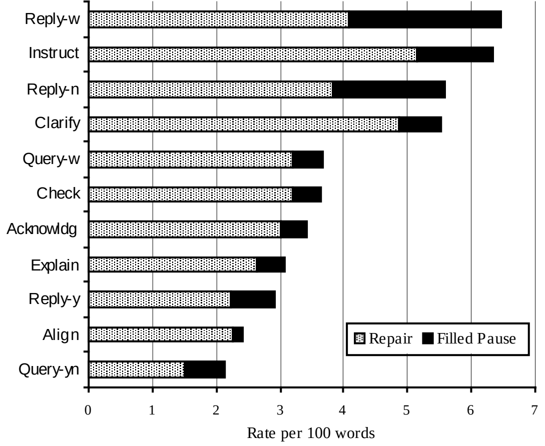


Figure 3 shows overall disfluency rates, split by repair and filled pause for each move-type. Amongst analyses for rates of different types of disfluency for this subset of the data, all four types of *replies to questions* form part of the most disfluent homogeneous subset (of 5 types) for the rate of *Filled Pauses*. **Reply-w** (2.4) and **reply-n** (1.78) are at the top of the group, with **reply-y** (0.68) and **clarify** (0.65) following **instruct** moves (1.18). **Clarify** and **reply-w** moves also show the highest rates for *disfluent repetition*, and for the full data set, they form their own subset for this dependent variable.

Instruct moves stand out amongst initiate-type moves as having high disfluency rates of all types.

4. Discussion

Clearly, dialogue moves differ in the extent to which they are likely to contain disfluencies, and longer moves tend to be more disfluent than shorter moves. But even when we level out move-length, significant differences in disfluency rates of all types remain between move-types.

Instructions in a route-giving domain entail planning, creativity and introducing new referents and typically involve hesitation and self-correction more than many other types of move.

Answering questions requires time, and speakers appear to use filled pauses and repetitions in order to gain time. The more difficult it is to find and formulate an answer, the more disfluency speakers produce: positive replies (**reply-y**) in the MTC typically confirm the presence of a landmark on the map, which involves less searching than affirming absence (**reply-n**); more complex replies (**reply-w**, **clarify**) may require time to find the answer as well as providing more opportunity for error.

We can begin to explain why instruction givers in the MTC and in [8] are more disfluent than followers. The act of constructing an instruction produces a high rate of disfluencies, the **instruct** move accounts for about 30% of all words in the corpus and is almost exclusively a Giver move (99.2%). **Clarify** moves are also highly disfluent and mostly produced by Givers (94.9%).

We can also see why some corpora have lower disfluency rates than others. In simpler dialogues, like typical human-computer interactions, less scope is allowed for producing complex replies and longer utterances in general are suppressed (they also contain little, if any, speaker overlap, which may contribute to repetition and deletion rates).

5. Acknowledgments

This work was supported by the ESRC main grant to HCRC and by EPSRC Project Grant GR/L50280/01. Thanks to David McKelvie and Amy Isard for help at various stages of data handling and to Matthew Bull and Cathy Sotillo for assistance with disfluency annotation.

6. References

[1] Smith, V. and Clark, H. H., “On the course of answering questions”, *Journal of Memory and Language*, 32:25-38.

[2] Levelt, W.J.M., “Monitoring and self-repair in speech”, *Cognition*, 14:14-104, 1983.

[3] Postma, A. and Kolk, H., “The covert repair hypothesis: prearticulatory repair processes in normal and stuttered disfluencies”, *J. Speech Hearing Res.*, 36:472-487, 1993.

[4] Clark, H. H., “Using Language”, Cambridge University Press, Cambridge, 1996.

[5] Pickering, M. and Garrod, S., “Towards a mechanistic psychology of dialogue”, Manuscript in preparation, 2001.

[6] Schegloff, E. A., Jefferson, G. and Sacks, H., “The preference for self-correction in the organisation of repair in conversation”, *Language*, 53:361-382, 1977.

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

[8] Bortfeld, H., Leon, S.D., Bloom, J. E., Schober, M. F. and Brennan, S. E., “Disfluency rates in conversation: Effects of age, relationship, topic, role and gender”, *Language and Speech*, 44, 2001.

[9] Bard, E. G., Lickley, R. J. and Aylett, M. P., “Is disfluency just difficulty?”, *Proceedings of Disfluency in Spontaneous Speech ’01, ISCA Tutorial and Research Workshop*, Edinburgh, Scotland, 2001.

[10] Lickley, R. J., “Detecting Disfluency in Spontaneous Speech”, PhD Thesis, University of Edinburgh, UK, 1994.

[11] Shriberg, E. E., “Preliminaries to a theory of speech disfluencies”, PhD Thesis, University of California at Berkeley, 1994.

[12] Brennan, S. E. and Williams, M., “The feeling of another’s knowing: Prosody and filled pauses as cues to listeners about the metacognitive state of speakers”, *Journal of Memory and Language*, 34:393-398, 1995.

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

[14] Greene, J. O. and Capella, J. N., “Cognition and talk: the relationship of semantic units to temporal patterns of fluency in spontaneous speech”, *Language and Speech*, 29(2):141-157, 1986.

[15] Oviatt, S., “Predicting disfluencies during human-computer interaction”, *Computer Speech and Language*, 9:19-35, 1995.

[16] Anderson, A., Bader, M., Bard, E.G., Boyle, E., Doherty, G., et al., “The HCRC Map Task Corpus”, *Language and Speech*, 34:351-366, 1991.

[17] Isard, A., “An XML architecture for the HCRC Map Task Corpus”, *Proceedings of Bi-Dialog, 2001*, Bielefeld, Germany, 2001.

[18] Carletta, J. C., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., Anderson, A., “HCRC Dialogue structure coding manual”, *HCRC/TR-82*, HCRC, University of Edinburgh 1996.

[19] Carletta, J. C., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., Anderson, A., “The reliability of a dialogue structure coding scheme”, *Computational Linguistics,* 23:13-31, 1997.

[20] Lickley, R. J., “HCRC Disfluency Coding Manual” *HCRC/TR-100*, HCRC, University of Edinburgh, 1998.

[21] LTG, “LTXML”. http://www.ltg.ed.ac.uk/software/xml.