NON-LINGUISTIC INFLUENCES ON RATES OF DISFLUENCY IN SPONTANEOUS SPEECH

Holly Branigan[*], Robin Lickley[†] and David McKelvie[†]
[*]*Human Communication Research Centre, University of Glasgow, UK*
[†]*Human Communication Research Centre, University of Edinburgh, UK*

ABSTRACT

We investigate how non-linguistic factors influence rates of disfluency in spontaneous speech in a set of task-oriented dialogues (the HCRC Map Task Corpus). The factors we consider are: sex of the speaker; sex of the addressee; conversational role; ability to see the addressee; familiarity with the addressee; and practice at the task. Our analyses examined disfluency rate (the number of disfluencies per 100 intended words) and discard rate (the number of reparandum words per 100 intended words) in a series of within- and between-speaker comparisons. Our results suggest that, perhaps unsurprisingly, these non-linguistic factors do influence speaker fluency. However, their influences may manifest themselves through complex interactions with other factors, and in some cases may only be apparent for particular measures of disfluency or particular types of disfluency. The work has implications for both psychological and computational models of speech production and perception.

1. INTRODUCTION

Speaking spontaneously entails on-line planning, error-correction and interaction with other speakers. As a result, speakers often pause, backtrack to alter things that they have just said or abandon sentences midstream. There is no complete account of the causes of such disfluency, but it appears likely that non-linguistic factors may play an important part. In this paper, we examine rates of disfluency in the HCRC Map Task Corpus [1] and the extent to which these rates are affected by non-linguistic factors in the design of the corpus.

1.1. The Corpus

The Map Task Corpus consists of 128 dialogues (approx. 130,000 words in total) between pairs of Glasgow University students (32 males, 32 females). Dialogues averaged 407.75 seconds in duration and 1193 words. In each dialogue, both participants had a map which had various landmarks drawn on it. One participant (the Giver) also had a route marked on their map. Their task was to describe the route to the other participant (the Follower), who had to draw this route onto their own map. Givers averaged 820.27 words per dialogue and Followers 372.74. To make the task more complicated, the two maps had different, but overlapping, sets of landmarks.

The 64 speakers were divided into 16 groups of 4 (quads). In 8 quads, the participants had eye contact, but they could not see each other's maps; in the other 8 quads, participants were screened from each other. Within each quad were two pairs of friends; members of each pair were unfamiliar with members of the other pair. Each speaker took part in 4 dialogues, twice as Giver and twice as Follower. For each role, they took part once with a partner with whom they were familiar.

The design of the corpus thus allowed us to look at the effects of the following factors on rates of disfluency: **sex** of speaker and of addressee; **conversational role** (Giver vs. Follower); **eye contact**; familiarity with addressee; **practice** (first vs. second attempt at describing a route using the same map). The corpus allows us to look for consistent patterns of disfluency in a large number of speakers producing spontaneous task-oriented speech.

1.2. Disfluency coding

Disfluency labelling on the entire corpus was done by hand, using Xwaves/Entropic signal processing software and xlabel. The coder was able to see the speech waveform and examine a spectrogram where necessary, as well as being able to hear and replay the signal as many times as required. Labels were aligned with units of the word-level transcription, which was in the same format. Each disfluency was labelled for type, denoting the 'editing operation' (**Basic types**: repeat, delete, insert, substitute; **Non-basic types:** combinations of these and complex types involving embedding) and the number of words in the **reparandum** (the words which need to be removed to render a fluent interpretation). Filled pauses were coded separately but are not included in the following analyses. The coding method is described in detail in [2] and is similar to that used by Shriberg in [3].

All the coding was converted into XML, from which the data for the following study was extracted. **Disfluency rate** was calculated as number of disfluencies per 100 *intended* words (i.e. excluding filled pauses, editing terms and words in reparanda). Note that disfluency rate does not always give an accurate picture of the proportion of speech that is disfluent, since it does not take into account the length (in words) of each disfluency. We therefore also calculated **discard rate** (number of reparandum words per 100 intended words). Note that two speakers with comparable disfluency rates may differ greatly in the proportion of their speech that is disfluent.

1.3. Predictions

1.3.1. Sex. There is no strong independent motivation to lead us to predict that one sex should be more disfluent than the other. Lickley [4] found males to be more disfluent than females, but his corpus only contained 6 speakers. Shriberg [3] found that males used filled pauses more often than females (for

30 speakers in the Switchboard corpus). The Map Task Corpus allows us to explore for a larger number of speakers whether male speakers are more disfluent than females. It also lets us examine whether a speaker is more disfluent when talking to a same-sex or opposite-sex partner.

1.3.3. Role. The Map Task gives us examples of speech by the same speakers taking different roles in dialogues. As an instruction giver, the speaker has to decide how to describe a certain route across a map, circumventing various objects. This involves both complex conceptual planning, for example the order in which to tackle various problems, and complex linguistic planning, for example formulating an instruction involving multiple entities. In contrast, instruction followers have less of a planning task and less need to introduce new concepts to the dialogue. For the most part, their contribution consists of acknowledging that they have understood their instructions. Both conceptual and linguistic planning load might plausibly be expected to influence disfluency rates. Because Givers in general have the heavier planning load, we would therefore predict that Givers should produce more disfluencies.

However, there is also a possible confound between role and utterance length. Givers tend to produce longer utterances than Followers. Some previous research has suggested that longer utterances have higher disfluency rates than shorter utterances [5]. Hence Givers might have higher disfluency rates simply because they tend to produce longer utterances.

1.3.3. Familiarity. As stated above, each speaker took part in dialogues with one partner with whom they were familiar, as well as with a partner with whom they were not familiar. Familiarity could affect disfluency rates in at least two ways. On the one hand, speakers might be more careful and cooperative when speaking to an unfamiliar addressee. This could manifest itself in a greater propensity to pre-plan an utterance before beginning to speak, resulting in lower disfluency rates with unfamiliar than familiar addressees. Alternatively, speakers might be less anxious when speaking to a familiar addressee, and it is possible that this might result in lower disfluency rates with familiar addressees. (Note that we have no independent means of determining whether speakers' anxiety levels were any different in the two conditions.) Familiarity with the other speaker's speech style and turn-taking tactics might also lead to fewer disfluencies in dialogues between familiar participants.

1.3.4. Eye Contact. Speakers took part in dialogues in one of two Eye-contact conditions. Bull [6] found that eye contact had an effect on inter-turn intervals: Where speakers had eye contact, the average time between turns was longer than when they did not have eye contact. It may be that eye signals are useful for effective turn-trading. This could result in lower disfluency rates in the Eye-contact condition than the No eye-contact condition for two reasons. First, speakers might be more aware of when the other person wishes to speak. Second, speakers might be better able to detect when the addressee is having difficulty, and hence might be able to help the addressee out (e.g. by clarifying an instruction) without the addressee having to interrupt. Both factors would reduce the number of abandoned utterances caused by the addressee interrupting.

1.3.5. Practice. Each speaker performed the role of Giver twice with the same map. Practice with the task should mean that a Giver has less conceptual planning to do on the second attempt and this should lead to less hesitation (fewer repetitions). There may also be fewer disfluencies caused by Followers interrupting in the second session: There may be less need for the Follower to interrupt if the Giver remembers where the maps differ and can therefore accommodate the Follower.

2. RESULTS

2.1. Overview

Table 1 shows the disfluency rate and discard rate for the corpus as a whole, averaged over speakers.

	Mean	Min	Max	S D
Disfluency Rate	4.05	1.55	9.48	1.49
Discard Rate	8.20	2.68	18.07	3.18

Table 1: Disfluency rates and discard rates per 100 fluent words.

There is considerable inter-speaker variation for both measures. But unsurprisingly, there was a significant correlation by speaker between disfluency rate and discard rate (Pearson's product moment correlation: $r = .955$, $p < .01$).

The overall rate for each type of disfluency and the percentage of all disfluencies accounted for by each type is shown in Table 2. (Non-basic types (4.74% of all disfluencies) are excluded from the table.)

	Measure	Mean	Min	Max	S D
Reptn	Disfl rate	1.28	0.28	5.24	0.79
	% all disfl	30.54	15.63	60.87	0.10
Deltn	Disfl rate	1.67	0.56	3.38	0.62
	% all disfl	42.22	21.26	65.31	0.10
Instn	Disfl rate	0.40	0.05	1.05	0.23
	% all disfl	9.70	1.61	21.26	0.04
Subst	Disfl rate	0.51	0.00	1.18	0.23
	% all disfl	12.80	0.00	31.58	0.05

Table 2: Disfluency rates per 100 fluent words, and percentage of all disfluencies, by type.

Speakers varied considerably both in disfluency rate by type, and in the proportion of each type of disfluency that they produced. This was particularly noticeable for repetitions and deletions, which together constituted on average 73% of all disfluencies. There was a positive correlation between an individual's overall repetition and deletion rates (Pearson's correlation: $r = .303$, $p < .05$). Thus individuals with higher repetition rates also tended to have higher deletion rates: Some individuals are consistently more disfluent than others.

However, speakers varied in the relative proportions of repetitions and deletions that they produced. For 28 speakers (more than one third of the total) the difference between the proportion of repetitions and deletions produced was more than 20%. These results are in keeping with Shriberg's [3] finding of individual 'styles' of disfluency. However, our data do not provide clear evidence of a bimodal distribution of 'Repeaters' and 'Deleters'. Rather, individual speakers' preferences for repetition versus deletion appear to vary along a continuum.

2.2. Influence of non-linguistic factors

2.2.1. Sex. Tables 3 and 4 show average disfluency rates and discard rates respectively for male and female speakers. We report rates for the corpus as a whole and for the two Eye-contact conditions individually.

	Cond	Mean	Min	Max	S D
Male	Overall	4.35	1.84	9.48	1.62

	Eye-c	4.19	2.51	5.49	0.97
	No eye-c	4.52	1.84	9.48	2.12
Fmle	Overall	3.76	1.55	6.37	1.42
	Eye-c	3.16	1.55	5.85	1.35
	No eye-c	3.90	1.98	5.79	1.33

Table 3: Average disfluency rates for males and females, for entire corpus, eye-contact, and no eye-contact conditions.

	Cond	Mean	Min	Max	S D
Male	Overall	8.57	2.68	18.07	3.39
	Eye-c	8.56	6.19	12.08	2.05
	No eye-c	8.57	2.68	18.07	4.48
Fmle	Overall	7.70	3.16	13.63	3.07
	Eye-c	6.49	3.16	11.97	2.77
	No eye-c	7.80	3.57	12.24	3.12

Table 4: Average discard rates for males and females, for entire corpus, eye-contact, and no eye-contact conditions.

Although males were numerically more disfluent than females on both measures, the difference was only marginally significant for disfluency rate and non-significant for discard rate (Independent-pairs t-tests, one-tailed: $t(62) = 1.604$, $p = .06$; $t(62) = .211$, $p > .1$).

Further tests examined data from the Eye-contact and No eye-contact conditions separately. These showed that in the Eye-contact conditions, females had significantly lower disfluency and discard rates than males ($t(20) = 2.052$, $p < .05$; $t(20) = 1.994$, $p < .05$, both one-tailed). However, there were no such differences in the No eye-contact conditions (Both $t < 1$, $p > .1$, one-tailed). One interpretation of these results is that females may be better than males at picking up visual cues for effective turn-trading from their partners.

Further tests examined whether speakers' disfluency and discard rates were influenced by their partner's sex. Numerically, disfluency rates and discard rates were higher with same-sex partners than opposite-sex partners (4.48 vs. 4.31; 9.48 vs. 8.67), but these differences were not reliable ($t(27) = .705$, $p > .1$; $t(27) = 1.345$, $p > .1$, both two-tailed). The results suggest that there is an overall difference between the sexes in disfluency and discard rates, but there is no influence of the sex of the addressee.

2.2.2. Role. Table 5 shows average disfluency rates and discard rates for Givers and Followers. Speakers had higher disfluency and discard rates as Givers than as Followers (Paired-samples t-tests, both one-tailed: $t(63) = 4.846$, $p < .01$; $t(63) = 5.547$, $p < .01$). Further tests (all one-tailed) compared disfluency rates for each disfluency type. These revealed that Givers were reliably more disfluent than Followers with respect to repetitions ($t(63) = 2.471$, $p < .05$); insertions ($t(63) = 6.552$, $p < .01$); substitutions ($t(63) = 5.552$, $p < .01$); and marginally more disfluent with respect to deletions ($t(63) = 1.469$, $p = .08$). These differences can plausibly be attributed to the more complex conceptual and linguistic processing that Givers must carry out than Followers.

	Measure	Mean	Min	Max	S D
Giver	Disfl rate	4.30	1.56	11.37	1.70
	Discard rate	8.83	2.79	22.12	3.72
Follower	Disfl rate	3.40	0.80	7.20	1.54
	Discard rate	6.57	1.91	13.62	3.14

Table 5: Average disfluency rates and discard rates for Givers and Followers.

We also analysed the relationship between utterance length and disfluency rate. On average, 71% of Followers' utterances were fewer than 5 words long, compared to 49% for Givers. The higher a Follower's proportion of utterances under 5 words, the lower their disfluency and discard rates (Pearson's product moment correlations: $r = -.437$, $p < .01$; $r = -.446$, $p < .01$). However, there were no such correlations for Givers (both $p > .1$). Thus, Followers' but not Givers' disfluency rates were predictable from their proportion of short utterances.

Givers and Followers also produced differing proportions of each type of disfluency. Proportionally more of Givers' disfluencies than Followers' were insertions and substitutions (Paired-samples t-tests, one-tailed: $t(63) = 4.673$, $p < .01$; $t(63) = 3.302$, $p < .01$); whereas more of Followers' disfluencies than Givers' were deletions ($t(63) = 3.058$, $p < .01$). There was no difference for repetitions ($p > .1$).

Overall, these results suggest that conversational role has a considerable influence upon disfluency rate and type of disfluency. The different in rate cannot be wholly attributed to differences in utterance length. This is an important issue for future investigation.

2.2.3. Familiarity. Table 6 shows average disfluency rates and discard rates for Givers with familiar and unfamiliar Followers.

	Measure	Mean	Min	Max	S D
Familiar Follower	Disfl rate	4.34	1.21	11.28	1.87
	Discard rate	9.09	1.50	21.54	4.37
Unfam Follower	Disfl rate	4.12	1.29	11.44	1.81
	Discard rate	8.22	1.50	22.57	3.76

Table 7: Average disfluency rates and discard rates for Givers with familiar and unfamiliar Follower.

For both measures, rates were numerically higher with familiar than unfamiliar Followers. The difference was not significant for disfluency rate (Paired-samples t-test: $p > .1$, 2-tailed). However, Givers had significantly higher discard rates, i.e. a higher proportion of words appearing in reparanda, when speaking to a familiar Follower ($t(63) = 2.139$, $p < .05$, 2-tailed).

2.2.4. Eye-contact. Table 7 shows average disfluency rates and discard rates in the Eye-contact and No eye-contact conditions. Both disfluency and discard rates were numerically higher for the No eye-contact condition. However, this difference was not reliable (Independent-samples t-tests, one-tailed: both $p > .1$).

	Measure	Mean	Min	Max	S D
Eye-c	Disfl rate	3.88	1.55	7.43	1.30
	Discard rate	7.99	3.16	15.86	2.80
No eye-c	Disfl rate	4.23	1.84	9.48	1.66
	Discard rate	8.42	2.68	18.07	3.55

Table 7: Average disfluency rates and discard rates in Eye-contact and No eye-contact conditions.

Analyses investigating each type of disfluency separately revealed higher rates of repetition disfluencies in the No eye-contact condition ($t(62) = 1.952$, $p < .05$, one-tailed). This may be attributable to more effective turn-trading in the Eye-contact condition, resulting in less need for repetition in an attempt to gain the floor. There was no comparable effect for other types of disfluency (all $p > .1$).

2.2.5. Practice. Table 8 shows Givers' average disfluency rates and discard rates in their first and second sessions as Giver.

	Measure	Mean	Min	Max	S D
1st turn	Disfl rate	4.35	1.29	11.44	1.85
	Discard rate	8.62	1.50	22.57	4.23
2nd turn	Disfl rate	4.13	1.21	11.28	1.80
	Discard rate	8.75	1.69	21.54	3.88

Table 8: Givers' average disfluency rates and discard rates in first and second sessions as Giver.

There was no reliable difference in overall disfluency and discard rates between Givers' first and second sessions (Paired-sample t-tests, one-tailed: $t(63) = 1.317$, $p = .1$; $t(63) = .306$, $p > .1$). However, analyses investigating each type of disfluency separately revealed higher rates of repetition disfluencies in speakers' first session as Giver ($t(63) = 2.136$, $p < .05$, one-tailed). There was no comparable effect for other types of disfluency (all $p > .1$). Repetitions appear to be primarily hesitation devices, whereas other types of disfluency tend to reflect alterations to the current speech plan. The finding of lower rates of repetition with practice is therefore compatible with the hypothesis that practice at a task results in less hesitation, due to lower conceptual planning demands.

3. CONCLUSIONS

Our results show that speakers' fluency in spontaneous speech is influenced by a number of non-linguistic factors:

- Female speakers were less disfluent than males under certain circumstances: when they could see their addressee. Sex of the addressee did not affect fluency.
- Instruction givers were more disfluent than instruction followers; this difference does not appear to be solely attributable to differences in utterance length.
- Conversational role also affects the types of disfluencies produced: Repetition rates did not differ between roles but Givers produced proportionally more insertions and substitutions, and fewer deletions, than Followers.
- Speakers had a higher discard rate (proportionally more words forming part of a reparandum) when speaking to a familiar addressee. But familiarity did not affect the disfluency rate (i.e. the *number* of disfluencies that speakers produced).
- Speakers produced more repetitions when they could not see their addressee. But eye contact did not reliably affect overall disfluency or discard rate.
- Instruction givers produced fewer repetitions the second time that they performed the task. But practice did not influence overall disfluency or discard rate.

Our results show further that non-linguistic factors do not exert a uniform effect on fluency. Many differences, though showing the predicted pattern numerically, did not achieve statistical significance. In addition, some analyses achieved significance for one measure only (e.g. disfluency rate but not discard rate or vice versa). This indicates that it may be important to employ more than one measure of disfluency in order to uncover underlying patterns in the data.

More interestingly, the results suggest that different factors may interact in complex ways. Thus the speaker's ability to see the listener does not in itself significantly affect rates of disfluency; but an interaction of eye contact and speaker's sex turns out to exert a strong influence on both disfluency and discard rates. Hence, an important conclusion of this work is that it may be over-simplistic to expect simple relationships between non-linguistic factors and disfluency. Equally, our results suggest that it may be simplistic to concentrate on gross measures of disfluency, and that analyses which distinguish between different categories of disfluency may ultimately be more enlightening.

Future work will examine the interactions more closely. We will also add other hesitation phenomena (filled and silent pauses) to the analyses and take into account possible effects of speech rate, inter-turn intervals, and utterance length and type.

ACKNOWLEDGMENTS

We thank Amy Isard for technical help, Matthew Bull for coding assistance and Ellen Gurman Bard for useful advice. The first author was supported by a British Academy Postdoctoral Fellowship and acknowledges a BT Short Term Research Fellowship. The second and third authors were supported by EPSRC SALT grant number GR/L50280.

REFERENCES

[1] Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H. & Weinert, R. 1991 The HCRC Map Task Corpus. *Language and Speech*, 34, 351-366.

[2] Lickley, R.J., 1998. HCRC Disfluency Coding Manual. *HCRC Technical Report* 100.

[3] Shriberg, E.E., 1994. Preliminaries to a theory of speech disfluencies. PhD Thesis. UC Berkeley.

[4] Lickley, R.J. 1994. Detecting disfluency in spontaneous speech. PhD Thesis, University of Edinburgh.

[5] Oviatt, S. 1995. Predicting spoken disfluencies during human-computer interaction. *Computer Speech and Language* 9, 9-15.

[6] Bull, M. 1998. The timing and coordination of turn-taking. PhD Thesis, University of Edinburgh.