DISFLUENCY IN MULTIPARTY SOCIAL TALK

Emer Gilmartin, Carl Vogel, Nick Campbell
Trinity College Dublin
{gilmare, vogel, nick}@tcd.ie

Keywords: disfluency, hesitation, repair, casual conversation, spoken interaction

1. INTRODUCTION

Much research on disfluencies in spontaneous spoken interaction has been carried out on corpora of task-based conversations, resulting in greater understanding of the role of several phenomena. Modern multimodal corpora allow the full spectrum of signals in face to face communication to be analysed. However, the 'unmarked' case of casual conversation or social talk with no obvious short-term instrumental goal has been less studied in this manner. Corpus-based work on social talk tends to deal with short dyadic interactions, although the norm for social conversation is for longer multiparty interaction. In this paper, we outline our programme of exploratory studies of disfluency in a longer multiparty conversation. We briefly describe the background to our research goals, and then report on the collection, transcription, and annotation of the data for our experiments. We present and discuss some of our early results.

2. SOCIAL TALK

Casual social conversation is defined as talk engaged in when 'talking just for the sake of talking'[7], and includes smalltalk, gossip, and conversational narrative. Malinowski described 'phatic communion' as an emergent activity of congregating people, comprising free aimless social conversation, which he viewed as the most basic use of speech [15]. For Malinowski, the purpose of such talk is not to exchange information or to express thought, but to avoid unfriendly silence and strengthen social bonding. This view is echoed in Jakobson's phatic component in his model of communication [11], Brown and Yule's distinction between interactional and instrumental language [4], and Dunbar's theory that language evolved to maintain social cohesion through verbal grooming as group size grew too large for physical grooming [6]. Laver focuses on the 'psychologically crucial margins of interaction', conversational openings and closings, postulating that small talk performs a lubricating or transitional function from silence to greetings to business and back to closing sequences and to leave taking [13]. Schneider analysed audio recordings of naturally occurring small talk, concentrating on the linguistic content of entire dialogues [21]. He described instances of small talk at several levels, from frames such as 'WEATHER' to sequences and adjacency pairs, and their constituent utterance types. He identifies idling sequences of repetitions of agreeing tails such as 'Yes, of course', 'MmHmm' as prevalent in social talk. Ventola viewed casual conversation as composed of several phases - with ritualised opening greetings, followed by approach segments of light uncontroversial small talk, which sometimes led to longer and more informative centre phases consisting of sequential but overlapping topics, and then back to ritualised leavetakings [23]. Thus a social conversation could range from a simple exchange of greetings, through a short exchange of small talk, to longer more varied stretches of spoken interaction covering several topics. Slade and Eggins state that through casual conversation people form and refine their social reality [7]. They cite gossip, where participants reaffirm their solidarity by jointly ascribing outsider status to another, and show examples of conversation between friends at a dinner party where greater intimacy allows differences of opinion. They identify story-telling as a frequent genre in conversation and highlight 'chat' (interactive exchanges involving short turns by all participants) and 'chunks' (longer uninterrupted contributions) elements of conversation. They report that casual conversation tends to involve multiple participants rather than the dyads normally found in instrumental interactions or examples from conversation analysis. Instrumental and interactional exchanges differ in duration; task-based conversations are bounded by task completion and tend to be short, while casual conversation can go on indefinitely. Indeed, early conversation analysts pointed out that these casual conversational situations or 'continuing state(s) of incipient talk' were not covered by the theories of (task-based) conversational structure being developed [19].

It seems likely that the distribution of disfluencies, both within turn pauses, hesitations and repairs, and phenomena such as recycled restarts and abandoned utterances, will vary between different types

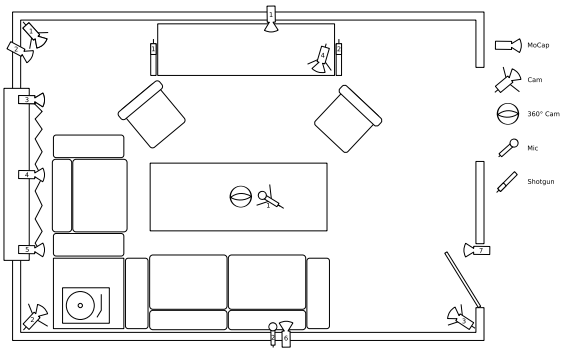
of interaction, and across different phases or subgenres of the same interaction. If, as is frequently claimed, disfluencies are a mark of planning difficulties or cognitive load, they should be less frequent in shorter more ritualised small talk or approach sequences. In more central sequences, they may appear more often in discussion sequences or at the beginning of narrative sequences as the speaker 'gets going', but less in idling sequences. Distributions may also vary between social and task-based conversations. In order to test these ideas, we have prepared a 70-minute sample of extended social talk data, on which we are currently experimenting, as described below.

3. DATA AND ANNOTATION

Corpora used for studies of disfluencies in human-human non-pathological adult spontaneous speech in English include the HCRC MapTask corpus of dyadic information gap task-based conversations [1], SWITCHBOARD corpus of dyadic telephonic speech [9], ICSI and AMI multiparty meeting corpora [12] [17], and resources such as recordings of televised political interviews [2]. However, the speech in these resources, while spontaneous and conversational, does not meet our need for longer multiparty social face to face conversation data. Therefore, for our preliminary studies, we have prepared a sample drawn from a longer conversation in the D64 corpus of spontaneous multiparty social talk.

The D64 corpus is a multimodal corpus of over 8 hours of informal conversational English recorded in an apartment living room, as shown in Fig. 1. Several streams of video, audio, and motion capture data were recorded for the corpus. There were between 2 and 5 people on camera at all times. There were no instructions to participants about what to talk about and care was taken to ensure that all participants understood that they were free to talk or not as the mood took them. Design and collection of the corpus is fully described in[18]. The audio recordings included near-field chest or adjacent microphone recordings for each speaker. These were found to be unsuitable for automatic segmentation as there were frequent overlaps and bleedover from other speakers. After a manual synchronisation, the audio files for each speaker were segmented manually into speech and silence intervals using Praat [3] on 10 and 4-second or smaller windows as necessary. The process was then repeated for the sound file recorded at the same time for each of the other speakers, resulting in annotations checked across five different sound files. Any remaining speech intervals not assigned to a particular speaker were resolved using Elan [24] to refer to the video recordings taken at the same time. There are concerns to note with the manual segmentation into speech and silence and indeed in manual annotation of disfluencies, as human hearing and comprehension is a filter rather than a simple sensor. Humans listening to speech can miss or imagine the existence of objectively measured silences of short duration, especially when there is elongation of previous or following syllables [16], and are known to have difficulty recalling disfluencies from audio they have heard [5]. However, in the current work, speech can be slowed down and replayed and, by zooming in on the waveform and spectrogram, annotators can clearly see silences and differences in amplitude on the speech waveform and spectrogram. This need to match the heard linguistic and non-linguistic content to the viewed waveform and spectrogram means that it was much more likely that pauses and disfluencies would be noticed.

Figure 1: Setup for Session 1 of D64 Recordings.



After segmentation the data for Session 1 of the corpus were manually transcribed and annotated, using a scheme largely derived from the TRAINS transcription scheme [10]. Words, hesitations, filled and unfilled pauses, unfinished words, laughs and coughs were transcribed and marked. The transcription was carried out at the intonational phrase (IP) level rather than the more commonly used interpausal unit (IPU) as IPs are a basic unit for intonation study and can easily be concatentated to the interpausal unit (IPU) and turn level as required. The transcriptions were then text-processed for automatic word alignment, which was carried by running the Penn Aligner [25] over a sound file and accompanying transcription for each intonational phrase annotated. Sections which could not be automat-

ically aligned, where there was significant overlap or cut off words, were manually aligned. In order to more fully investigate genre within casual talk, conversational sections were labeled as discussion, dominated, or idling. Idling was labelled orthogonally to discussion and dominated as it could occur within either modality. Discussion referred to stretches of talk shared more or less evenly among two or more participants throughout the bout, while dominated referred to bouts largely dominated by one participant. These often took the form of narratives or recounts of personal experiences, extended explanations or opinions. A total of 142 'bouts' were annotated, of which 14 were labelled as 'discussion' while the remaining 128 were classed as dominated.

Table 1: The annotation code used for disfluencies.

Symbol	Note
.	interruption point
-	unfinished word
tilde	unfinished utterance
caret	contracted word
r	repeated word
s	substituted word
d	deleted word
f	filled pause
x	pause
o	overlap

The word level transcription was then used with the sound files to annotate disfluencies using Praat. The scheme and procedures used were based largely those outlined in Shriberg's and Eklund's respective theses [22] [8], and in Lickley's annotation manual for the MapTask corpus [14], with extra labels and coventions for recycled turn beginnings [20], disfluencies in the presence of overlapping speech from another participant, and unfinished and abandoned utterances. The symbols used are outlined in Table. 1. Complex, or nested, disfluencies were labelled following Shriberg's method [22], and no indexing was used for substitutions or repetitions. Pauses within utterances were annotated with 'x' when they occurred within a larger disfluency or with '[.x]' when they occurred alone. The annotated data sample comprised 15,545 words across 6164 intonational phrase units, with 1505 annotated disfluencies. There were 653 lone pauses, which were removed from the dataset for the purposes of this analysis. Of the remaining 853 disfluencies, 117 were complex. Just over 15%, 128 disfluencies, occurred in the presence of overlap by another speaker

Below we describe preliminary results on the distribution of disfluencies in the corpus in general and particularly disfluencies in the presence or absence of overlapping speech in the dominated genre.

4. RESULTS AND CONCLUSIONS

We concentrate on the dominated genre of talk, comprising 777 disfluencies. For preliminary investigation, complex disfluencies were removed from the dataset, leaving 668 disfluencies, of which 13%, 87 disfluencies, were in the presence of overlap.

Table 2: Disfluency types in overlap in dominated genre (%).

Sp	Del	Rep	Sub	FP
All	52	34	3	11
Main	59	18	5	18
Other	67	25	2	6

In the overlap condition, the distribution of disfluency type over all disfluencies, and for the dominant or main speaker and other speakers in the dominated genre is shown in Table. 2 . It can be seen that the distributions are similar for both the main and other speakers with the bulk of disfluencies in overlap occurring as deletions of unfinished utterances. This finding is consistent across both dominant speakers and other speakers. However, speakers other than the main speaker were even more likely to abandon their utterance, while the main speaker used more filled pauses in overlap, possibly to indicate intention to continue.

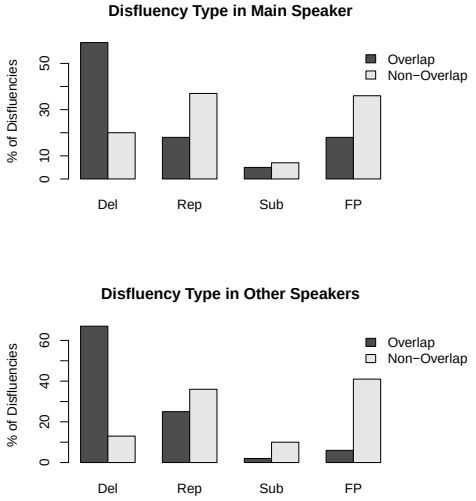
Table 3: Disfluency types in non-overlap in dominated genre (%).

Sp	Del	Rep	Sub	FP
All	18	36	8	38
Main	20	37	7	36
Other	13	36	10	41

In the non-overlap condition there were 581 disfluencies, distributions for all speakers, main and other speakers can be seen in Table, 3. In this condition, deletions and filled pauses are the most common types of disfluency and almost equally common, with the distribution showing consistency for main and other speakers. Figure. 2 contrasts the frequency of each disfluency type in main and other speakers in the overlap and non-overlap environments.

Our preliminary results show a strong tendency for speakers to stop in the presence of overlap,

Figure 2: Disfluency distributions in overlap and non-overlap environments.



although this does not always happen. It would be very interesting to analyse whether this tendency and indeed the occurrence of disfluency varies throughout the course of a bout of dominated conversation for main and other speakers. However, we cannot simply use distance of disfluencies from the bout start as the proportion of speech by different participants varies as each bout progresses. We are currently working on a measure of distance from the start of each bout which takes account of this variation. This will allow us to further explore the role of disfluency in casual multiparty conversation.

5. REFERENCES

[1] Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., others, 1991. The HCRC map task corpus. *Language and speech* 34(4), 351–366.

[2] Beattie, G. 1983. *Talk: An analysis of speech and non-verbal behaviour in conversation*. Open University Press.

[3] Boersma, P., Weenink, D. 2010. *Praat: doing phonetics by computer [Computer program], Version 5.1. 44*.

[4] Brown, G., Yule, G. 1983. *Teaching the spoken language* volume 2. Cambridge University Press.

[5] Deese, J. 1980. *Pauses, prosody, and the demands of production in language*. Mouton Publishers.

[6] Dunbar, R. 1998. *Grooming, gossip, and the evolution of language*. Harvard Univ Press.

[7] Eggins, S., Slade, D. 2004. *Analysing casual conversation*. Equinox Publishing Ltd.

[8] Eklund, R. 2004. *Disfluency in Swedish human-human and human-machine travel booking dialogues*. LinkkÌoping, Sweden: Dept. of Computer and Information Science, Linköping Studies in Science and Technology.

[9] Godfrey, J. J., Holliman, E. C., McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on* volume 1 517–520.

[10] Heeman, P. A., Allen, J. F. 1995. The TRAINS 93 Dialogues. Technical report DTIC Document.

[11] Jakobson, R. 1960. Closing statement: Linguistics and poetics. *Style in language* 350, 377.

[12] Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A. 2003. The ICSI meeting corpus. *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03). 2003 IEEE International Conference on* volume 1 I–364.

[13] Laver, J. 1975. Communicative functions of phatic communion. *Organization of behavior in face-to-face interaction* 215–238.

[14] Lickley, R. J. 1998. HCRC disfluency coding manual. *Human Communication Research Centre, University of Edinburgh*.

[15] Malinowski, B. 1923. The problem of meaning in primitive languages. *Supplementary in the Meaning of Meaning* 1–84.

[16] Martin, J. G. 1970. On judging pauses in spontaneous speech. *Journal of Verbal Learning and Verbal Behavior* 9(1), 75–78.

[17] McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V. 2005. The AMI meeting corpus. *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research* volume 88.

[18] Oertel, C., Cummins, F., Edlund, J., Wagner, P., Campbell, N. 2010. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces* 1–10.

[19] Schegloff, E., Sacks, H. 1973. Opening up closings. *Semiotica* 8(4), 289–327.

[20] Schegloff, E. A. 1987. Recycled turn beginnings: A precise repair mechanism in conversation's turn-taking organization. *Talk and social organization* 70–85.

[21] Schneider, K. P. 1988. *Small talk: Analysing phatic discourse* volume 1. Hitzeroth Marburg.

[22] Shriberg, E. E. 1994. *Preliminaries to a theory of speech disfluencies*. PhD thesis University of California.

[23] Ventola, E. 1979. The structure of casual conversation in English. *Journal of Pragmatics* 3(3), 267–298.

[24] Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. 2006. Elan: a professional framework for multimodality research. *Proceedings of LREC* volume 2006.

[25] Yuan, J., Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123(5), 3878.