Exploring Multiparty Casual Talk for Social Human-Machine Dialogue

Emer Gilmartin(✉), Benjamin R. Cowan, Carl Vogel, and Nick Campbell

Trinity College, Dublin, Ireland
{gilmare,vogel,nick}@tcd.ie, benjamin.cowan@ucd.ie

Abstract. Much talk between humans is casual and multiparty. It facilitates social bonding and mutual co-presence rather than strictly being used to exchange information in order to complete well-defined practical tasks. Artificial partners that are capable of participating as a speaker or listener in such talk would be useful for companionship, educational, and social contexts. However, such applications require dialogue structure beyond simple question/answer routines. While there is body of theory on multiparty casual talk, there is a lack of work quantifying such phenomena. This is critical if we are to manage and generate human machine multiparty casual talk. We outline the current knowledge on the structure of casual talk, describe our investigations in this domain, summarise our findings on timing, laughter, and disfluency in this domain, and discuss how they can inform the design and implementation of truly social machine dialogue partners.

Keywords: Speech interfaces · Dialogue modelling · Casual social talk

1 Introduction

Human talk is a fundamentally social activity, and casual conversation is inevitable whenever humans gather together. It forms a fundamental part of human communication. With the growth of interest in the development of avatars and robots as social companions, it is important to understand the nature of such talk in situations where there is more than one conversational actor so as to endow machines with the ability to converse appropriately in such contexts. Currently, much of the speech interface research is focused on task based dialogue interactions. Early dialogue system researchers recognised the complexity of dealing with social talk [1], and initial prototypes concentrated on practical tasks such as travel bookings and the logistics of moving boxcars of oranges. In these tasks, the lexical content of utterances was enough to drive successful completion of the task. Task-based systems have proven invaluable in many practical domains. However the desire to develop more social companions (be it robot or avatar based) in a number of domains such as healthcare and education means that social talk must become a significant strand of research and the nature of dialogue as a multi-party activity needs to be addressed. We argue that the field

A. Karpov et al. (Eds.): SPECOM 2017, LNAI 10458, pp. 370–378, 2017.
DOI: 10.1007/978-3-319-66429-3_36

now needs to move towards understanding and incorporating casual multiparty conversation so as to create more natural dialogue interactions between machine and human partners. In this paper we highlight work in the area of social talk and summarise recent research conducted by the authors in this domain.

2 What Is Social Talk?

Social talk, rather than following Gricean maxims of efficient communication of information, seems rather to be based on avoidance of silence and engagement in unthreatening but entertaining verbal display and interaction [32]. In casual talk, all participants can contribute at any time, unlike the more restricted roles found in more formal situations [14,34]. Rather than following a question and answer format of the type which drives task based dialogues, casual conversation has been described as occurring in stages - chat and chunk [17]. In chat phases, participants contribute utterances more or less equally with many questions and short comments. Chat is often used to 'break the ice' among strangers involved in casual talk [28]. Chat phases are also interspersed with chunk phases – longer contributions from one participant – often in the form of narrative anecdotes and recounts, opinion or discussion. The 'ownership' of chunks seems to pass around the participants in the talk, with chat linking one chunk to the next [17]. The structure of casual conversation has also been described as a more detailed sequence of structural elements or phases [33]. These phases include opening and closing Greeting, Address, Leavetaking and Goodbye sequences. The main content of the conversation is described as a sequence of Approach and Centring stages, similar to chat and chunk, with added subtypes for the Approach phases depending on social distance between participants. Approach phases can be indirect - dealing with topics such as the weather, or direct - involving more personal subject matter. Figure 1 shows a schematic of the phases described by Ventola, while Fig. 2 shows examples drawn from our data of typical chat and chunk phases in a 5-party conversation.

The design of more social speech interfaces and companion applications depends on knowledge of the type of talk being modelled. Below we outline our work in this area, focussing on corpus analysis to determine the characteristics of longer form multiparty casual talk.

3 Corpora Used for Casual Conversation Research

Relevant corpora of human interaction are essential to understanding different genres of spoken dialogue. Interestingly, the design of systems and the production of corpora has often followed the path taken in the development of pragmatic theories of talk. Early task-based systems were based on a literal view of speech as transmission of text. Many of the multimodal corpora and indeed several earlier audio corpora created in laboratory and 'real-world' conditions have been collections of performances of the same spoken task by different subjects, or of interactions specific to particular domains where lexical content was fundamental

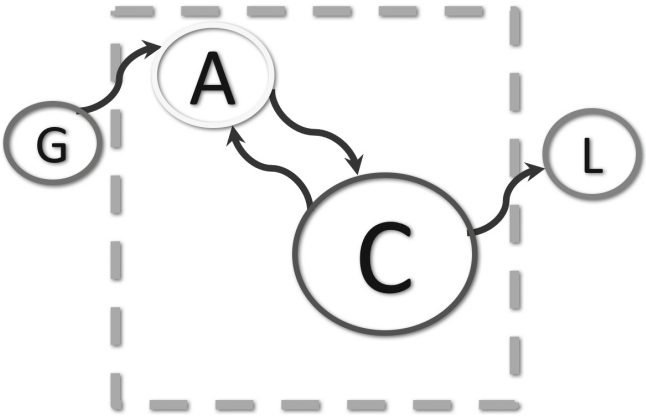

Fig. 1. The phases of casual talk described by Ventola - greeting, approach, centre, and leavetaking. Note that approach and centre phases may freely recur

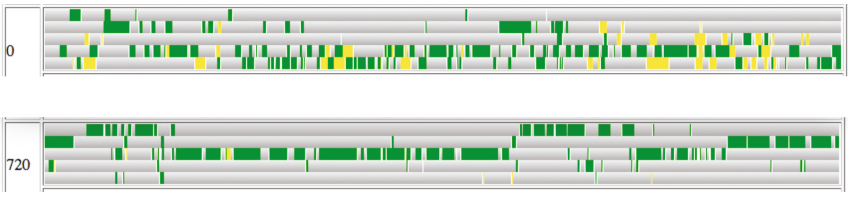

Fig. 2. Examples of chat (top) and chunk (bottom) phases in two stretches from a 5-party conversation in the D64 corpus. each row denotes the activity of one speaker across 120 s. Speech is green, and laughter is yellow on a grey background (silence). The chat frame, taken at the beginning of the conversation, can be seen to involve shorter contributions from all participants with frequent laughter. The chunk frame shows longer single speaker stretches (Color figure online)

to acheivment of a practical goal - such corpora include collections of information gap activities such as the HCRC MapTask corpus of dyadic information gap task-based conversations [3]. Other corpora have focussed on collecting recordings of real or staged meetings, such as the ICSI and AMI multiparty meeting corpora [25,29], or recordings of particular genres of interaction, such as televised political interviews [6]. All of these corpora have contributed greatly to understanding of different facets of spoken interaction such as timing, turntaking, and dialogue architecture. However, the speech in these resources, while spontaneous and conversational, cannot be considered casual talk, and the results obtained from their analysis may not transfer to casual conversation.

In terms of non-task interaction, there have been audio collections made of casual talk, including telephonic corpora such as SWITCHBOARD [22] and the ESP-C collection of Japanese telephone conversations [13], and corpora comprising recordings of face-to-face talk as in the Santa Barbara Corpus [15],

and sections of the ICE corpora [23] and of the British National Corpus [8]. These corpora are audio only and thus cannot be used to inform research on facial expression, gestural or postural research. The Gothenburg Corpus of recordings of different types of human activity contains both audio and video recordings including casual or small talk [2], leading a trend toward multimodal recordings which can be used to study more aspects of conversation.

Increasing interest in social talk among dialogue system designers has resulted in systems which engage users in 'chat' similar to the smalltalk described at the margins of more serious practical talk in the pragmatics literature [7,36]. In the recent years, researchers have started to produce corpora of mostly dyadic 'first encounters' where strangers were recorded engaged in casual conversation for periods of 5 to 20 min or so [4,16,31]. These corpora have appeared in several languages including Swedish, Danish, Finnish, and English. These corpora are very valuable for the study of dyadic interaction, particularly at the opening and early stages of interaction. For a fuller review of available corpora and the challenges of genre in conversation, see [18]. However, pragmatic work has described the substance of longer casual conversation beyond these first encounters, and it is this area which interests us, informing the design of systems which can take the user into a longitudinal series of conversations beyond the first chat phases.

We focus on multiparty casual conversation, and have created a dataset of six informal conversations with three to five participants, each around an hour long. The conversations were drawn from three multimodal corpora, d64, DANS, and TableTalk [12,24,30], to allow for comparison of our results from analysis of the audio data with results of video analysis at a later date. Recordings of this type are not easily found with those corpora being the most popular for such work. Our data was manually segmented and transcribed using Praat [9] and Elan [35]. Details of the dataset can be seen in Table 1, and further details of the annotation process can be found in [20]. In the next section, we give an overview of recent work on this dataset.

Table 1. Source corpora and details for the conversations used in dataset

Corpus	Participants	Gender	Duration (s)
D64	5	2F/3M	4164
DANS	3	1F/2M	4672
DANS	4	1F/3M	4378
DANS	3	2F/1M	3004
TableTalk	4	2F/2M	2072
TableTalk	5	3F/2M	4740

In each of the corpora used, participants were recorded in casual conversation in a living room setting or around a table, with no instructions on topic of type of conversation to be carried out - participants were also clearly informed that they could speak or stay silent as the mood took them.

4 Overview of Recent Work

Our analysis of social talk focuses on a number of dimensions; chat and chunk duration, laughter distribution, disfluency distribution, and the patterning of utterances by different speakers in different phases, as these elements are largely independent of the lexical content of the conversations, and have been analysed in meeting corpora [5,11,27]. Thus, our analyses of casual multiparty talk can be contrasted with existing analyses of task-based multiparty talk. Timing information in multiparty meeting corpora, in particular, has been shown to be amenable to stochastic modelling of the distribution of talk and laughter [26], which is a longer term goal of the work described here.

4.1 Chat and Chunk Duration and Chat Positioning

From our analysis of data from the corpora highlighted we have found that the distributions of durations of chat and chunk phases are different, with chat phases durations varying more while chunk durations have a more consistent clustering around the mean. Chat phase durations (Mean = 28 s) tend to be shorter than chunk durations (Mean = 34 s). These findings are not speaker specific in our preliminary experiments and seem to indicate a natural limit for the time one speaker should dominate a conversation. The dimensions of chat and chunk durations observed would indicate that social talk should 'dose' or package information to fit chat and chunk segments of roughly these lengths. In particular, the tendency towards chunks of around half a minute could help in the design of narrative or education-delivering speech applications, by allowing designers to partition content optimally.

We also observed more chat at conversation beginnings, with chat predominating for the first 8–10 min of conversations. Although our sample size is small, this observation conforms to descriptions of casual talk in the literature, and reflects the structure of 'first encounter' recordings. However, as the conversation develops, chunks start to occur much more frequently, and the structure is an alternation of single-speaker chunks interleaved with shorter chat segments. While the initial extended chat segments can be used to model 'getting to know you' sessions, and will therefore be useful for familiarisation with a digital companion, it is clear that we need to model the chunk heavy central segments of conversation if we want to create systems which form a longer-term dialogic relationship with users.

4.2 Laughter and Disfluency Distribution

We have also been investigating the frequency and distribution of laughter and disfluencies in multiparty casual talk. Early experiments showed that laughter, and particularly shared laughter, appears more common in social talk than in meeting data, and that laughter happens more around topic endings/topic changes [10,19]. This is consistent with our current work on chat and chunk phases, as we are seeing that laughter is more common in chat phases – which

provide a 'buffer' between single speaker and topic chunks. In the current dataset we have found that laughter accounts for approximately 10% of vocal time in chat phases while it only accounts for 4% of chunk phases. For disfluencies, a pilot study has shown differences in the occurrence and distribution of disfluency types for chunk owners in chunks and all other speakers [21]. In the chunk modality one speaker holds the floor for an extended period and this behaviour is different to that of all other speakers in chunks, to that of all speakers in chat, and indeed to that of the chunk owner when in somebody else's chunk.

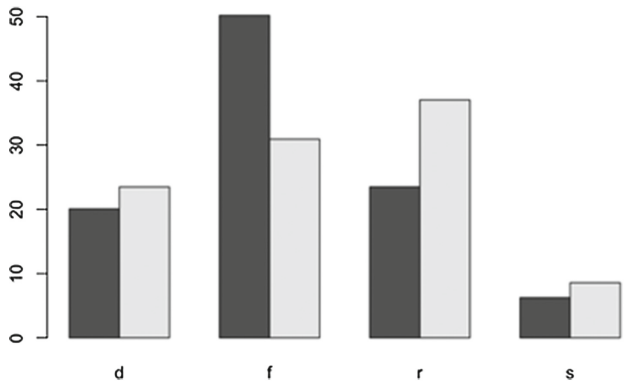

Fig. 3. Distribution of disfluency types (deletion, filled pause, repetition, substitution) in chunk owner versus all other speech. Frequencies are shown proportionally in percentages with grey denoting chunk owner speech.

Figure 3 shows the distribution of disfluency types (deletion, filled pause, repetition, substitution) in two modalities – where the speaker is the? owner? of a chunk versus all other speech. It can be seen that filled pauses are proportionally less frequent in chunk owner speech than in general speech – 31% vs 50%, while repetition is proportionately more common in chunk owner speech – 37% vs 23%. In view of the very small sample of speakers, we checked the distributions for each speaker, although the proportions varied. For individual speakers, in all cases, filled pauses were also proportionately lower in chunk owner speech versus other speech, and repetitions were also proportionally higher in chunk owner speech for each speaker.

4.3 Speaker Contribution

We are studying the patterning of speaker contributions in both phases, particularly the length of gap or overlap in the vicinity of speaker and phase changes. We are performing prosodic analysis of the utterance final pitch movements in different contexts, and believe the results of this work will provide information helpful in developing more finegrained 'endpointing' systems to determine *when*

the system should speak; with knowledge of how turntaking occurs in different phases of talk we can work towards providing systems with turntaking behaviour appropriate to the current conversational phase.

5 Systems Developed for Casual Talk

Based on our analysis we have built a number of prototype 'first encounter' systems whose purpose is to chat engagingly with users. The HERME robot, based on casual talk structure, successfully chatted with several hundred members of the public in Trinity College's Science Gallery. Our more recent system, CARA, has been used in Wizard of Oz experiments to investigate timing by humans versus automatic machine timing in first encounter dialogues. We are currently developing CARA as a system which will incorporate our growing knowledge of how longer form casual talk actually works.

6 Conclusions

There is increasing interest in academic circles, business, and from the general public in spoken dialogue systems that act naturally and perform functions beyond information search and narrow task-based exchanges. The design of these new systems needs to be informed by relevant data and analysis of human spoken interaction in the domains of interest. Many of the available multiparty data are based on meetings or first encounters. While first encounters are very relevant to the design of human machine first encounters, there is a lack of data on longer human conversations. We hope that the encouraging results of our analysis of casual social talk will help make the case for the creation and analysis of corpora of longer social dialogues. We believe that the exponential growth in speech technology and companion systems means that data and scientific investigation around this type of talk is urgently needed so as to design more effective automated dialogue partners.

Acknowledgements. This work is supported by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences and Technologies ERA-NET (CHISTERA) JOKER project, JOKe and Empathy of a Robot/ECA: Towards social and affective relations with a robot, and by the Speech Communication Lab, Trinity College Dublin, and by Science Foundation Ireland funding for ADAPT (13/RC/2106) at Trinity College Dublin.

References

1. Allen, J., Byron, D., Dzikovska, M., Ferguson, G., Galescu, L., Stent, A.: An architecture for a generic dialogue shell. Nat. Lang. Eng. **6**(3&4), 213–228 (2000)
2. Allwood, J., Björnberg, M., Grönqvist, L., Ahlsén, E., Ottesjö, C.: The spoken language corpus at the department of linguistics, Göteborg University. In: FQS-Forum Qualitative Social Research, vol. 1 (2000). http://www.ling.gu.se/~jens/publications/bfiles/B45.pdf

3. Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al.: The HCRC map task corpus. Lang. Speech **34**(4), 351–366 (1991)
4. Aubrey, A.J., Marshall, D., Rosin, P.L., Vandeventer, J., Cunningham, D.W., Wallraven, C.: Cardiff conversation database (CCDb): a database of natural dyadic conversations. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 277–282, June 2013
5. Baron, D., Shriberg, E., Stolcke, A.: Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. Channels **20**(61), 41 (2002)
6. Beattie, G.: Talk: an Analysis of Speech and Non-verbal Behaviour in Conversation. Open University Press, Milton Keynes (1983)
7. Bickmore, T., Cassell, J.: Social dialogue with embodied conversational agents. In: van Kuppevelt, J., Dybkjaer, L., Bernsen, N. (eds.) Advances in Natural Multimodal Dialogue Systems, pp. 23–54. Kluwer, New York (2005)
8. BNC-Consortium: British national corpus (2000). URL http://www.hcu.ox.ac.uk/BNC
9. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [Computer program], Version 5.1. 44 (2010)
10. Bonin, F., Campbell, N., Vogel, C.: Laughter and topic changes: temporal distribution and information flow. In: 2012 IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom), pp. 53–58 (2012)
11. Bonin, F., Campbell, N., Vogel, C.: Temporal distribution of laughter in conversation. In: Proceedings of the Third Interdisciplinary Workshop on Laughter and other Non-Verbal Vocalization in Speech, pp. 25–26 (2012)
12. Campbell, N.: Multimodal processing of discourse information, the effect of synchrony. In: Second International Symposium on Universal Communication, ISUC 2008, pp. 12–15 (2008)
13. Campbell, N.: Approaches to conversational speech rhythm: speech activity in two-person telephone dialogues. In: Proceedings of XVIth International Congress of the Phonetic Sciences, Saarbrucken, Germany, pp. 343–348 (2007). http://www.icphs2007.de/conference/Papers/1775/1775.pdf
14. Cheepen, C.: The Predictability of Informal Conversation. Pinter, London (1988)
15. DuBois, J.W., Chafe, W.L., Meyer, C., Thompson, S.A.: Santa Barbara corpus of spoken american english. Linguistic Data Consortium, CD-ROM, Philadelphia (2000)
16. Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., House, D.: Spontal: a swedish spontaneous dialogue corpus of audio, video and motion capture. In: LREC (2010)
17. Eggins, S., Slade, D.: Analysing Casual Conversation. Equinox Publishing Ltd., Sheffield (2004)
18. Gilmartin, E., Bonin, F., Cerrato, L., Vogel, C., Campbell, N.: What's the game and who's got the ball? genre in spoken interaction. In: 2015 AAAI Spring Symposium Series (2015)
19. Gilmartin, E., Bonin, F., Vogel, C., Campbell, N.: Laugher and topic transition in multiparty conversation. In: Proceedings of the SIGDIAL 2013 Conference, pp. 304–308. Association for Computational Linguistics, Metz, August 2013
20. Gilmartin, E., Campbell, N.: Capturing chat: annotation and tools for multiparty casual conversation. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016) (2016)

21. Gilmartin, E., Vogel, C., Campbell, N.: Disfluency in multiparty social talk. In: Proceedings of DISS 2015, Edinburgh (2015)
22. Godfrey, J.J., Holliman, E.C., McDaniel, J.: SWITCHBOARD: telephone speech corpus for research and development. In: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-92, vol. 1, pp. 517–520 (1992)
23. Greenbaum, S.: ICE: the international corpus of english. Engl. Today **28**(7.4), 3–7 (1991)
24. Hennig, S., Chellali, R., Campbell, N.: The D-ANS corpus: the Dublin-Autonomous Nervous System corpus of biosignal and multimodal recordings of conversational speech. Reykjavik, Iceland (2014)
25. Janin, A., Baron, D., Edwards, J., Ellis, D., Gelbart, D., Morgan, N., Peskin, B., Pfau, T., Shriberg, E., Stolcke, A.: The ICSI meeting corpus. In: 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, Proceedings, ICASSP 2003, vol. 1, pp. I-364 (2003)
26. Laskowski, K.: Predicting, detecting and explaining the occurrence of vocal activity in multi-party conversation. Ph.D. thesis, Carnegie Mellon University (2011)
27. Laskowski, K., Burger, S.: Analysis of the occurrence of laughter in meetings. In: INTERSPEECH, pp. 1258–1261 (2007)
28. Laver, J.: Communicative functions of phatic communion. In: Organization of Behavior in Face-to-Face Interaction, pp. 215–238 (1975)
29. McCowan, I., Carletta, J., Kraaij, W., Ashby, S., Bourban, S., Flynn, M., Guillemot, M., Hain, T., Kadlec, J., Karaiskos, V.: The AMI meeting corpus. In: Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research. vol. 88 (2005)
30. Oertel, C., Cummins, F., Edlund, J., Wagner, P., Campbell, N.: D64: a corpus of richly recorded conversational interaction. J. Multimodal User Interfaces **7**, 1–10 (2010)
31. Paggio, P., Allwood, J., Ahlsén, E., Jokinen, K.: The NOMCO multimodal Nordic resource-goals and characteristics (2010). http://bada.hb.se/handle/2320/7400
32. Schneider, K.P.: Small Talk: Analysing Phatic Discourse, vol. 1. Hitzeroth, Marburg (1988)
33. Ventola, E.: The structure of casual conversation in English. J. Pragmatics **3**(3), 267–298 (1979)
34. Wilson, J.: On the Boundaries of Conversation, vol. 10. Pergamon Press, Pergamon (1989)
35. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: Elan: a professional framework for multimodality research. In: Proceedings of LREC, vol. 2006 (2006)
36. Yu, Z., Xu, Z., Black, A.W., Rudnicky, A.: Strategy and policy learning for non-task-oriented conversational systems. In: Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue (2016)