Capturing Chat: Annotation and Tools for Multiparty Casual Conversation.

Emer Gilmartin, Nick Campbell
Speech Communication Lab
Trinity College Dublin
gilmare@tcd.ie, nick@tcd.ie

Abstract

Casual multiparty conversation is an understudied but very common genre of spoken interaction, whose analysis presents a number of challenges in terms of data scarcity and annotation. We describe the annotation process used on the d64 and DANS multimodal corpora of multiparty casual talk, which have been manually segmented, transcribed, annotated for laughter and disfluencies, and aligned using the Penn Aligner. We also describe a visualization tool, STAVE, developed during the annotation process, which allows long stretches of talk or indeed entire conversations to be viewed, aiding preliminary identification of features and patterns worthy of analysis. It is hoped that this tool will be of use to other researchers working in this field.

Keywords: multimodal corpora, casual conversation, annotation

1. Introduction

Casual conversation is a fundamental feature of human social life, seemingly present in most situations where people congregate. In order to better understand the workings of casual conversation, suitable data are needed to inform studies of the various aspects of this most common of human activities. Although there are increasing numbers of high quality annotated multimodal corpora of spoken interaction available, there remain very few collections of longer stretches of casual talk among multiple participants. While it is tempting to use some of the rich resources currently available, it is unclear that results obtained from them would apply to structurally different casual talk. Therefore, we have undertaken a programme of collection, annotation, and analysis of multiparty casual talk data. In this paper, we outline the annotation process and tools developed during our ongoing programme of research into multiparty casual conversation, which aims to deepen understanding of this very common modality of spoken interaction, and inform the design of human-machine interaction. We briefly discuss models of social or casual conversation from the literature, provide an overview of corpora currently available and their limitations for our requirements, describe the annotation procedures we have developed in our work on two corpora of long (one hour or more) multiparty conversations, and introduce a visualisation tool developed to aid in preliminary examination and analysis of these conversations.

2. Social Talk

Casual social conversation is described as 'talking just for the sake of talking'(Eggins and Slade, 2004), and its subgenres include smalltalk, gossip, and conversational narrative. Aimless social talk or 'phatic communion' has been described as an emergent activity of congregating people, and viewed as the most basic use of speech (Malinowski, 1923). Researchers in fields including anthropology, evolutionary psychology, and communication have theorized that the such talk functions to build social bonds and avoid unfriendly or threatening silence, rather than simply to exchange information or express thought, as postulated in much linguistic theory. Instances of these views are found in the phatic component in Jakobson's model of communication (Jakobson, 1960), distinctions between interactional and instrumental language (Brown and Yule, 1983), and theories that language evolved to maintain social cohesion through verbal grooming (Dunbar, 1998). It has long been speculated that the prosodic and gestural aspects of social talk carry much of its communicative load, that 'how' things are said is as important as 'what' is said (Abercrombie, 1956; Hayakawa, 1990). Studies of casual conversation have focussed on the form and content of small talk and its discourse and sociolinguistic functions. Early analytic work focussed on the 'psychologically crucial margins of interaction', conversational openings and closings, with suggestions that small talk performs a lubricating or transitional function allowing talk to progress from initial silence through stages of greeting, business or 'meat' of the interaction, and back to closing sequences and to leave taking (Laver, 1975). The structure of casual conversation has been described in terms of distinct phases; often beginning with ritualised opening greetings, followed by approach segments of light uncontroversial small talk, and in longer conversations leading to more informative centre phases consisting of sequential but overlapping topics, and then back to ritualised leavetakings (Ventola, 1979). Schneider collected and analysed a corpus audio recordings of naturally occurring small talk, concentrating on the linguistic content of entire dialogues (Schneider, 1988). He described instances of small talk at several levels, from frames such as 'FOOD' to adjacency pairs and their constituent utterance types. Schneider also highlighted features prevalent in casual talk which did not seem to conform to Gricean ideas of dialogue - in particular, idling sequences of repetitions of agreeing tails such as 'Yes, of course', 'MmHmm' which seem to keep the conversation going rather than add any new information. He proposed a set of maxims peculiar to this genre, concentrated on the importance of avoiding silence and maintaining politeness, and suggested that Grice's Co-operative Principle itself remained relevant to small talk although several of the related maxims did not apply. Many researchers have high-

lighted the divide between the structure and characteristics of written and spoken language and have cited a possible text bias in linguistics as a retarding factor on the analysis of spoken interaction (Ong, 1982; Chafe and Danielewicz, 1987; Halliday, 1989). Syntactical, lexical, and discourse differences between (casual) conversation and more formal spoken and written genres are described in Biber and Leech's work on the Longman Corpus of Spoken and Written English (LSWE), and particularly in their chapter on the grammar of conversation (Biber et al., 1999). In terms of function, Slade and Eggins view casual conversation as the space in which people form and refine their social reality (Eggins and Slade, 2004) citing gossip between workmates, where participants reaffirm their solidarity, and examples of conversation between friends at a dinner party where greater intimacy allows differences of opinion. They identify story-telling as frequent in conversation and highlight segments of 'chat' (interactive exchanges involving short turns by all participants) and 'chunks' (longer uninterrupted contributions). They also 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. Several researchers on casual conversation have noted that their analyses were limited as they were based on transcripts and thus lacked vital timing and multimodal information.

3. Conversational Data

Much earlier work in conversation and discourse analysis was based on transcripts of audio recordings or indeed on written records of conversations heard, and thus timing information has not always been considered in great depth. Earlier data sources were audio only, and thus could not access the complete bundle of audio and video information now regarded as contributing to spoken interaction. With greater availability of recording equipment for modes including video, motion capture and indeed biosignals, researchers have been producing multimodal corpora which allow the full spectrum of signals in face to face communication to be analysed. 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. Corpora such as the HCRC MapTask corpus of dyadic information gap task-based conversations (Anderson et al., 1991), ICSI and AMI multiparty meeting corpora (Janin et al., 2003; McCowan et al., 2005), and resources such as recordings of televised political interviews (Beattie, 1983) have contributed greatly to our 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 the less studied 'unmarked' case of casual conversation. There have been audio collections made of casual talk, including telephonic corpora such as SWITCHBOARD (Godfrey et al., 1992) and the ESP-C collection of Japanese telephone conversations (Campbell, 2007), and corpora comprising recordings of face-to-face talk as in the Santa Barbara Corpus (DuBois et al., 2000), and sections of the ICE corpora (Greenbaum, 1991) and of the British National Corpus (BNC-Consortium, 2000). The Gothenburg Corpus of recordings of different types of human activity contains both audio and video recordings including casual or small talk (Allwood et al., 2000). Recently, multimodal corpora of spontaneous talk have been appearing in several languages. These include collections of free-talk meetings, or 'first encounters' between strangers as in the Swedish Spontal, and the NOMCO and MOMCO Danish and Maltese corpora (Edlund et al., 2010; Paggio et al., 2010). 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 (Gilmartin et al., 2015a))

Our focus is on longer stretches of face to face multiparty social talk, for which there are very few data collections available. In order to analyse social talk and ultimately create systems capable of performing or understanding this type of interaction, we have collected a number of recordings of casual multiparty speech. The multiparty nature of these interactions has created several challenges for annotation and analysis. Below we describe the workflow we have been developing for annotation and tools and resources we have generated.

4. Data and Annotation

The d64 corpus is a multimodal corpus of over 8 hours of informal conversational English recorded in Dublin in 2009 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(Oertel et al., 2010). The DANS corpus contains multimodal recordings of hour-long conversations between several participants recorded in a living-room setup in the Speech Communication Lab in Dublin in 2012. There are three participants on camera at all times. Again participants were encouraged to talk or not as the mood took them. The data include biosignal recorded using Affectiva Q-Sensors and heart-rate monitors. The design and collection of the corpus is described in (Hennig et al., 2014).

In each of the corpora audio recordings were made using near-field chest or adjacent microphone recordings for each speaker. The recordings were found to be unsuitable for automatic segmentation as there were frequent overlaps and bleedover from other speakers. While automatic segmentation could handle stretches where only one or two participants were talking without overlap, many turn changes involved overlap and there was significant choral production of short utterances and laughter. After manual synchronisation, the audio files for each speaker were segmented manu-

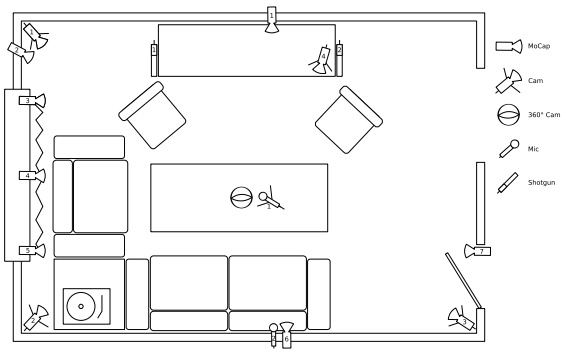

Figure 1: Setup for Session 1 of d64 Recordings.

ally into speech and silence intervals using Praat (Boersma and Weenink, 2010) on 10 and 4-second or smaller windows as necessary. The process was then repeated for the sound files recorded at the same time for each of the other speakers, resulting in annotations checked across several different sound files. Any remaining speech intervals or stretches of laughter or other vocalisations not assigned to a particular speaker were resolved using Elan (Wittenburg et al., 2006) to refer to the video recordings taken at the same time. There are valid concerns about manual segmentation into speech and silence, as human hearing and comprehension is a filter rather than a simple sensor. Thus, 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 (Martin, 1970), and are known to have difficulty recalling disfluencies from audio they have heard (Deese, 1980). 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. It is hoped that 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 and correctly marked.

After segmentation the data were manually transcribed and annotated, using a scheme largely derived from the TRAINS transcription scheme (Heeman and Allen, 1995). 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 and phoneme alignment with the Penn Aligner (Yuan and Liberman, 2008). The preparatory text-processing stage was accomplished using custom Praat and Python scripts to normalise the transcriptions and create an extension to the CMU dictionary used in alignment. The Penn Aligner was then run over a sound file and accompanying transcription for each intonational phrase annotated. Sections which could not be automatically aligned, where there was significant overlap or cut off words, were manually aligned. The word transcription was then corrected using Praat scripts to remove extra spaces added by the aligner. The segmentation, transcription, and alignment phases were performed on both d64 and DANS.

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

Table 1: The annotation code used for disfluencies.

The word level transcription was then used with the sound files to annotate disfluencies in Session 1 of the d64 coprus using Praat. The scheme and procedures used were based on those outlined in Shriberg's and Eklund's respective theses (Shriberg, 1994) (Eklund, 2004), and in Lickley's annotation manual for the MapTask corpus (Lickley, 1998), with extra labels and conventions for recycled turn beginnings (Schegloff, 1987), 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 (Shriberg, 1994), 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 fully annotated Session 1 of d64 comprised 15,545 words across 6164 intonational phrase units, with 1505 annotated disfluencies. There were 653 lone pauses. Of the remaining 853 disfluencies, 117 were complex. Just over 15%, 128 disfluencies, occurred in the presence of overlap by another speaker.

In order to more fully investigate the subgenres within casual talk, conversation in the first session of d64 was labeled as discussion, dominated, or idling. Idling, as defined by Schneider as discussed above, 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. Thus, discussion and dominated correspond to Slade and Eggins' concepts of chat and chunk as discussed above. Dominated stretches 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.

These subgenre and disfluency annotation phases are being extended to the remaining sections of d64 and to the DANS corpus.

5. Visualisation Tool - STAVE

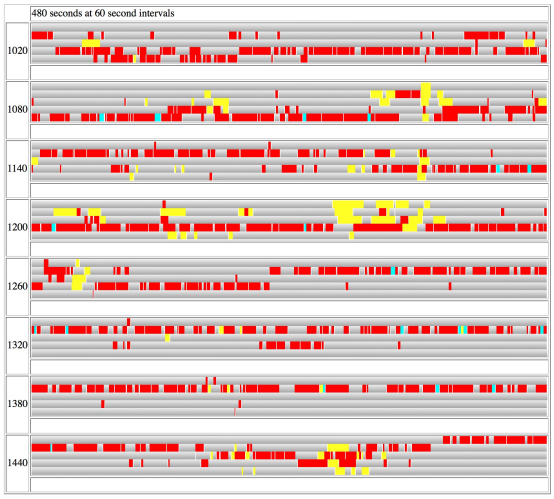

Figure 2: STAVE visualisation of speech (red), silence (grey), and laughter (yellow) in eight minutes from a five-party conversation.

For preliminary examination and analysis of possible trends in conversation data, we created a visualization tool, STAVE, which creates a visual HTML display representing a dialogue over time in the form of timelines for each speaker marked for speech, silence, laughter, disfluencies, or any other annotations desired. These are generated from a a comma-separated-variables (csv) file which can easily be obtained from any of the popular annotation software suites. The software, written in Python, can represent any time interval of conversation and colour codes can be assigned by the user. The software also generates colour-coded transcripts in Conversation Analysis format from a simple transcription file. It is possible to assign any combination of colours to different features of the annotation, including different colours per speaker tier.Sample STAVE output for eight minutes of the d64 corpus can be seen in Fig. 2. Data are arranged on a multi-tier timeline, in intervals chosen by the user (60 seconds in this case). Each grey line holds data for one speaker, showing speech (red), silence (grey), and laughter (yellow). This stretch is typical of the centre or 'steady-state' stage of conversation, with longer turns taken by each participant.

6. Conclusions and Future Work

We have described the segmentation, transcription and annotation of the d64 and DANS corpora of multiparty casual talk, and further subgenre and disfluency annotation of the first session of d64. The resulting annotations have already provided the basis for fruitful investigations into the architecture and features of this omnipresent but somewhat understudied form of spoken interaction - including experiments on laughter and disfluency (Gilmartin et al., 2013; Gilmartin et al., 2015b). We are currently studying the timing of speech and global silences in bouts of talk, and hope to use this knowledge to inform a novel timing module for a spoken dialogue system. Dialogue act annotation conforming to the ISO 24617-2 standard is underway on both corpora, using the ANVIL annotation tool and the DIAML annotation scheme, and the methodology outlined in (Bunt et al., 2012). These annotations will be included with future distributions of these corpora. We have also described visualisation tools designed for use with multispeaker recordings. All of these resources are being made available on the web, and it is hoped that they will be useful for other researchers.

7. Acknowledgements

This work is supported by the European Coordinated Research on Long-term Challenges in Information and Communication Sciences 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.

8. Bibliographical References

Abercrombie, D. (1956). *Problems and principles: Studies in the Teaching of English as a Second Language*. Longmans, Green.

Allwood, J., Björnberg, M., Grönqvist, L., Ahlsén, E., and Ottesjö, C. (2000). The spoken language corpus at the department of linguistics, Göteborg University. In *FQS–Forum Qualitative Social Research*, volume 1.

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

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

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., and Quirk, R. (1999). *Longman grammar of spoken and written English*, volume 2. Longman London.

BNC-Consortium. (2000). British national corpus. *URL http://www. hcu. ox. ac. uk/BNC*.

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

Brown, G. and Yule, G. (1983). *Teaching the spoken language*, volume 2. Cambridge University Press.

Bunt, H., Kipp, M., and Petukhova, V. (2012). Using DiAML and ANVIL for multimodal dialogue annotations. In *LREC*, pages 1301–1308.

Campbell, N. (2007). Approaches to conversational speech rhythm: Speech activity in two-person telephone dialogues. In *Proc XVIth International Congress of the Phonetic Sciences, Saarbrucken, Germany*, pages 343–348.

Chafe, W. and Danielewicz, J. (1987). *Properties of spoken and written language*. Academic Press.

Deese, J. (1980). *Pauses, prosody, and the demands of production in language*. Mouton Publishers.

DuBois, J. W., Chafe, W. L., Meyer, C., and Thompson, S. A. (2000). *Santa Barbara Corpus of Spoken American English. CD-ROM. Philadelphia: Linguistic Data Consortium*.

Dunbar, R. (1998). *Grooming, gossip, and the evolution of language*. Harvard Univ Press.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., and House, D. (2010). Spontal: A Swedish Spontaneous Dialogue Corpus of Audio, Video and Motion Capture. In *LREC*.

Eggins, S. and Slade, D. (2004). *Analysing casual conversation*. Equinox Publishing Ltd.

Eklund, R. (2004). Disfluency in Swedish human–human and human–machine travel booking dialogues.

Gilmartin, E., Hennig, S., Chellali, R., and Campbell, N. (2013). Exploring sounded and silent laughter in multiparty social interaction - audio, video and biometric signals. Valetta, Malta, October.

Gilmartin, E., Bonin, F., Cerrato, L., Vogel, C., and Campbell, N. (2015a). What's the Game and Who's Holding the Ball: Genre in Conversation. Palo Alto, California.

Gilmartin, E., Vogel, C., and Campbell, N. (2015b). Disfluency in Multiparty Social Talk. In *Proceedings of DISS 2015*, Edinburgh.

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

Greenbaum, S. (1991). ICE: The international corpus of English. *English Today*, 28(7.4):3–7.

Halliday, M. A. (1989). Spoken and written language.

Hayakawa, S. I. (1990). *Language in thought and action*. Houghton Mifflin Harcourt.

Heeman, P. A. and Allen, J. F. (1995). The TRAINS 93 Dialogues. Technical report, DTIC Document.

Hennig, S., Chellali, R., and Campbell, N. (2014). The D-ANS corpus: the Dublin-Autonomous Nervous System corpus of biosignal and multimodal recordings of conversational speech. Reykjavik, Iceland.

Jakobson, R. (1960). Closing statement: Linguistics and poetics. *Style in language*, 350:377.

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

Laver, J. (1975). Communicative functions of phatic communion. *Organization of behavior in face-to-face interaction*, pages 215–238.

Lickley, R. J. (1998). HCRC disfluency coding manual. *Human Communication Research Centre, University of Edinburgh*.

Malinowski, B. (1923). The problem of meaning in primitive languages. *Supplementary in the Meaning of Meaning*, pages 1–84.

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

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

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

Ong, W. J. (1982). Orality and literacy: The technology of the word. *New York: Methuen*.

Paggio, P., Allwood, J., Ahlsén, E., and Jokinen, K. (2010). The NOMCO multimodal Nordic resource–goals and characteristics.

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

Schneider, K. P. (1988). *Small talk: Analysing phatic discourse*, volume 1. Hitzeroth Marburg.

Shriberg, E. E. (1994). *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, University of California.

Ventola, E. (1979). The structure of casual conversation in English. *Journal of pragmatics*, 3(3):267–298.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., and Sloetjes, H. (2006). Elan: a professional framework for multimodality research. In *Proceedings of LREC*, volume 2006.

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