Composition and Dynamics of Multiparty Casual Conversation: A Corpus-based Analysis

Emer Gilmartin

Thesis submitted for the Degree of Doctor of Philosophy

School of Computer Science & Statistics

Trinity College

University of Dublin

2021

Declaration

I declare that this thesis has not been submitted as an exercise for a degree at this or any other university and it is entirely my own work. Wherever there is published or unpublished work included, it is duly acknowledged in the text.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

_____

Summary

This thesis provides a quantitative examination of multiparty casual talk, a speech genre which is fundamental to human social life. Multiparty casual conversation has been reported to comprise chat and chunk phases, where chat is interactive talk involving several participants and chunk is talk where one speaker dominates, holding the floor for a stretch of time. This thesis examines the current understanding of multiparty casual conversation, and reports several corpus studies performed in order to form a more accurate picture of its overall characteristics, and to compare and contrast its constituent chat and chunk phases in terms of composition and of conversation dynamics. This work aims to gain a quantitative understanding of the differences between chat and chunk phases, previously described by example, by analysis of speech, silence, laughter, conversation dynamics, and disfluency.

A collection of audio and video recordings of long (c. 1 hour) multiparty casual conversations (CasualTalk) was segmented, and annotated for speech, silence, laughter, and a subsection was annotated for disfluency. A novel chat and chunk annotation scheme was devised, tested, and applied to the corpus. The corpus was then segmented into novel floor state labels which reflected the conversation state at any instant in time.

The examination of the general structure of the casual conversations in CasualTalk supported accounts in the literature of such talk's structure of chat and chunk phases. The annotation process showed that chat and chunk phases can be annotated

with a high degree of inter-annotator agreement. Experiments on the compositions of chat and chunk phases showed that these do differ significantly in their distribution of speech and silence, proportion of laughter and overlap, and to a lesser extent in disfluency. For dynamics, there were clear differences in the liveliness of chat and chunk, both in terms of existing compression measures and a novel floor state change rate metric. The dynamics around turn change and retention also differed significantly. Analysis based on the identification of simple and complex within and between speaker transitions showed that more than half of all transitions are complex, and thus can be misclassified unless there is careful thresholding of the duration of single party speech bounding turns. Chat and chunk segments were compared to the task-based TEAMS Corpus, revealing significant differences in the distribution of silence and overlap, pointing to the need to consider and analyse different genres or speech exchange systems separately to form an accurate picture of their dynamics.

The results found will be useful in the design of more human-like dialog systems for a range of applications, particularly those requiring users to converse with the system in entertainment, education, or healthcare settings.

Acknowledgement

There are so many people I am grateful to for their help and company on the road to this document's final appearance that I'd need a second volume to mention everybody individually. So, first of all, I want to thank everyone who's discussed a question, shared ideas, collaborated on a project, co-organised a workshop, made a conference a venue for discovery and fun, or helped keep me sane over the last decade!

Special thanks are due to those who have guided my path in Trinity College Dublin. Ailbhe Ní Chasaide and Christer Gobl who first suggested the idea of my getting deeper into research and supported me throughout, Nick Campbell for guiding me through several years of this endeavour and introducing me to the wider world of speech technology, and especially to Carl Vogel, whose patience and kindness combined with intellectual rigour have kept me on track and interested through thick and thin. I also owe a debt of gratitude to Helen Thornbury, Niamh Farrelly and Claire Gleeson who helped immensely when I most needed support.

I've been a part of three different research labs in TCD - the Phonetics and Phonology lab, Speech Communications Lab, and the ADAPT Centre, each of which introduced me to a new group of researchers and friends. I have benefitted immensely from working (and playing!) with everybody here and with teams on a number of Enterface workshops over the years. A visit to Maxine Eskenazi at LTI in Pittsburgh near the end of this process gave me a new lease of life and reaffirmed my faith in research.

Outside the labs, I need to thank friends and family who have kept up the support

and encouragement through the years (and years...). Again, individual thanks would run to pages, so I'll just say without you I'd be lost.

And finally, thanks are due to my parents - my father, who instilled a love of language and learning in me and who has never failed to help, and to my mother, who always believed this day would come; although she is not here to see it, her loving kindness remains a guiding force.

Related Publications

F. Bonin, E. Gilmartin, C. Vogel, and N. Campbell. Topics for the future: Genre differentiation, annotation, and linguistic content integration in interaction analysis. In *Proceedings of the 2014 Workshop on Roadmapping the Future of Multimodal Interaction Research Including Business Opportunities and Challenges*, RFMIR '14, page 5–8, Istanbul, 2014. ACM.

H. Bunt, E. Gilmartin, S. Keizer, C. Pelachaud, V. Petukhova, L. Prévot, and M. Theune. Downward compatible revision of dialogue annotation. In *Proceedings of the 14th Joint ACL - ISO Workshop on Interoperable Semantic Annotation*, pages 21–34, Santa Fe, New Mexico, USA, Aug. 2018. Association for Computational Linguistics.

E. Gilmartin and N. Campbell. Capturing chat: Annotation and tools for multiparty casual conversation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4453–4457, Portoroz, Slovenia, 2016. European Language Resources Association (ELRA).

E. Gilmartin and C. Saam. Pragmatics research and non-task dialog technology. In *Proceedings of the 2nd Conference on Conversational User Interfaces*, page 1–3, Dublin, Ireland, 2020.

E. Gilmartin and C. Vogel. Chat, chunk and topic in casual conversation. In *Proceedings of the 14th Joint ACL-ISO Workshop on Interoperable Semantic Annotation*, page 45–52, Montpelier, France, 2018.

E. Gilmartin, F. Bonin, C. Vogel, and N. Campbell. Laugher and topic transition in multiparty conversation. In *Proceedings of the SIGDIAL 2013 Conference*, pages 304–308, Metz, France, Aug. 2013. Association for Computational Linguistics.

E. Gilmartin, F. Bonin, L. Cerrato, C. Vogel, and N. Campbell. What's the game and who's got the ball? genre in spoken interaction. In *Proceedings of the 2015 AAAI Spring Symposium Series*, Palo Alto, California, 2015a.

E. Gilmartin, C. Vogel, and N. Campbell. Disfluency in multiparty social talk. In *Proceedings of DISS 2015*, page 517–520, Edinburgh, Scotland, 2015b.

E. Gilmartin, B. R. Cowan, C. Vogel, and N. Campbell. Exploring multiparty casual talk for social human-machine dialogue. In A. Karpov, R. Potapova, and I. Mporas, editors, *Speech and Computer*, page 370–378. Springer International Publishing, 2017a.

E. Gilmartin, B. Spillane, M. O'Reilly, C. Saam, K. Su, B. R. Cowan, K. Levacher, A. C. Devesa, L. Cerrato, and N. Campbell. Annotation of greeting, introduction, and leavetaking in dialogues. In *Proceedings of the 13th joint ISO-ACL workshop on interoperable semantic annotation (ISA-13)*, Montpelier, France, 2017b.

E. Gilmartin, C. Vogel, and N. Campbell. Disfluency in chat and chunk phases of multiparty casual talk. In *Proceedings of DISS 2017*, page 25, Stockholm, Sweden, 2017c.

E. Gilmartin, M. O'Reilly, C. Saam, B. R. Cowan, C. Vogel, N. Campbell, and V. Wade. Silence and overlap in chat and chunk phases of multiparty casual conversation. In *Proceedings of the 9th International Conference on Speech Prosody 2018*, page 379–383, Poznan, Poland, 2018a.

E. Gilmartin, C. Saam, C. Vogel, N. Campbell, and V. Wade. Just talking-modelling ca-

sual conversation. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, page 51–59, Melbourne, Australia, 2018b.

E. Gilmartin, C. Vogel, and N. Campbell. Chats and chunks: Annotation and analysis of multiparty long casual conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1964–1970, Miyazaki, Japan, 2018c.

E. Gilmartin, B. R. Cowan, C. Vogel, and N. Campbell. Chunks in multiparty conversation—building blocks for extended social talk. In M. Eskenazi, L. Devillers, and J. Mariani, editors, *Advanced Social Interaction with Agents: 8th International Workshop on Spoken Dialog Systems*, Lecture Notes in Electrical Engineering, page 37–44. Springer International Publishing, 2019a.

E. Gilmartin, B. Spillane, C. Saam, C. Vogel, N. Campbell, and V. Wade. Stitching together the conversation—considerations in the design of extended social talk. In *Proceedings of the 9th International Workshop on Spoken Dialogue System Technology*, page 267–273, Syracuse, Italy, 2019b. Springer.

E. Gilmartin, M. Yu, and D. Litman. Comparing speech, silence and overlap dynamics in a task-based game and casual conversation. In *Proceedings of the 19th International Congress of Phonetic Sciences, 2019*, pages 3408–3412, Melbourne, Australia, 2019c.

E. Gilmartin, K. Aare, M. O'Reilly, and M. Włodarczak. Between and within speaker transitions in multiparty conversation. In *Proceedings of the 10th International Conference on Speech Prosody 2020*, 2020.

Contents

List of Tables

List of Figures

Part I

Background: Introduction and Literature Review

Chapter 1

Introduction

> *"What minimal model of the actor is needed if we are to wind him up, stick him in amongst his fellows, and have an orderly traffic of behavior emerge?"*
>
> - Goffman, 1967

1.1 Purpose of this Thesis

This thesis examines spontaneous multiparty casual conversation by analysing aspects of recorded multiparty conversations. The conversations of interest are casual in the sense that there are no predefined topics or practical task or transaction, such as a commercial exchange, to be accomplished. The speech in these conversations is spontaneous in the sense that it is not prepared in advance or scripted. Such talk is often termed social rather than task-based as it occurs between people with no clear reason other than propinquity, and no defined or mutually agreed purpose. The conversations of interest are multiparty in that they involve more than two participants.

Such conversations have been described as comprising a series of different phases – interactive 'chat' where several participants contribute more or less equally, and more monologic 'chunks' where one participant dominates the conversation. As with

much description of conversation, the definition is qualitative. The present analysis focusses on the distribution of speech, silence, laughter and disfluency, and the conditions around silence and overlap in multiparty casual conversations as a whole, and in chat and chunk phases of such conversations in particular. The goal is to add quantitative analysis of chat and chunk phases to better understand their composition in terms of these common features of conversation.

Although widely described as the fundamental form of spoken interaction, casual conversation has been somewhat overlooked in communication studies as a genre of human-human spoken interaction. Spoken interaction for a practical purpose, often dyadic, and indeed written language have received much more analysis, and are thus better described than casual talk. Research into casual conversation is now attracting interest with the exponential increase in (spoken) dialog applications which require more friendly or social artificial dialog. Earlier dialog technology concentrated on implementing dyadic task-based interactions, such as airline bookings, customer service responses, or pizza ordering. Recently, interest is shifting towards more 'friendly' systems which can perform tasks in a more human-like way, and also to applications where the system needs to build and maintain a longitudinal friendly relationship with one or more users - in therapeutic, educational, or companionship domains. A more accurate model of causal or social talk is essential for successfully modelling and engineering human-machine interaction and designing such systems.

The work in this thesis will ultimately contribute to the engineering goal of designing human-like dialogue systems which can engage in social talk with users in addition to performing task-based activities. The greater scientific goal is to better understand the mechanics of multiparty casual and social talk; this work will contribute to knowledge and understanding of the constituents of casual social spoken interaction, and thus to a broader understanding of human spoken interaction in all its facets. This thesis gathers existing knowledge, both theoretical and experimental,

of casual conversation in terms of its linguistic and paralinguistic forms and function, and reports new corpus studies on multiparty casual conversation data to extend understanding of the physical features of casual conversation – timing, structure, and the incidence of laughter and disfluency. It is hoped that this work will contribute to the construction of the minimal model of everyday human social behaviour alluded to by Goffman (Goffman, 1967, 3), adding to our understanding of this universal human activity and informing engineering applications using speech and dialog technology.

1.2 Methodology

Task-based talk has been studied quite extensively in recent years, due largely to the availability of data. Significant effort has gone into the collection of laboratory created corpora, where participants are recorded performing knowledge-gap tasks, and indeed of corpora of real or staged meetings. Casual conversational data is not as plentiful, particularly for English, and thus most descriptions of such talk are based on short examples transcribed or recalled by authors, or on opportunistic recordings collected in the past when ethics regulation was not as strict. More recent corpora of casual talk are often collections of short conversations, and do not provide adequate time for conversations to develop fully. The analysis in this thesis is based on CasualTalk - a set of recordings of multiparty long (c. 1hour) casual conversations in English recorded in living room like settings – a holiday apartment living room, around a table in a laboratory, and sofa and armchairs in a laboratory. In all of these interactions, participants were instructed to talk or not as the mood took them, and the conversations continued for up to 79 minutes.

Casual conversation proceeds as a series of contributions from participants, either speaking in the clear or in overlap. The pattern of who is speaking or not (the conversational floor state) changes constantly throughout a conversation. Analysis of

the distribution of speech, silence, and overlap should provide insight into the characteristics of conversations as a whole and of chat and chunk phases within those conversations. The incidence of laughter and disfluency overall and in chat and chunk phases should further elucidate the overall description of casual talk and the distinctions between chat and chunk phases. Research into the patterning of speech, silence, and laughter in conversation has generally followed two paths - large scale statistical modelling, and minute examination of examples of interaction. Large scale stochastic analysis uses very large data sets, machine learning and low level features such as speech and silence to infer and predict information about conversations, concentrating on the predictive power of models, rather than on building a human understandable picture of conversation. The latter paradigm, exemplified by Conversation Analysis, concentrates on the description and analysis of the organization of talk, using minute examination of meticulously transcribed examples from real conversations to highlight broader patterns of interaction noticed by the researchers, thus providing great detail on the local dynamics but not large scale statistical analysis. This work threads a middle path between very large scale prediction and fine grained explicative analysis in order to explore multiparty casual conversation, using corpus analysis of several hours of multiparty conversation to better understand the general characteristics of such talk, based on expectations from existing work on the topic.

1.3 Research goals and work performed for thesis

The thesis examines the current understanding of multiparty casual conversation, and performs corpus studies performed in order to form a more accurate picture of its overall characteristics, chat and chunk structure, and conversation dynamics. This general research goal can be broken down into a number of parts, with specific tasks and analsyses arising. Figure 1.1 provides an overview of the work performed.

The general research goal can be stated in three broad research questions

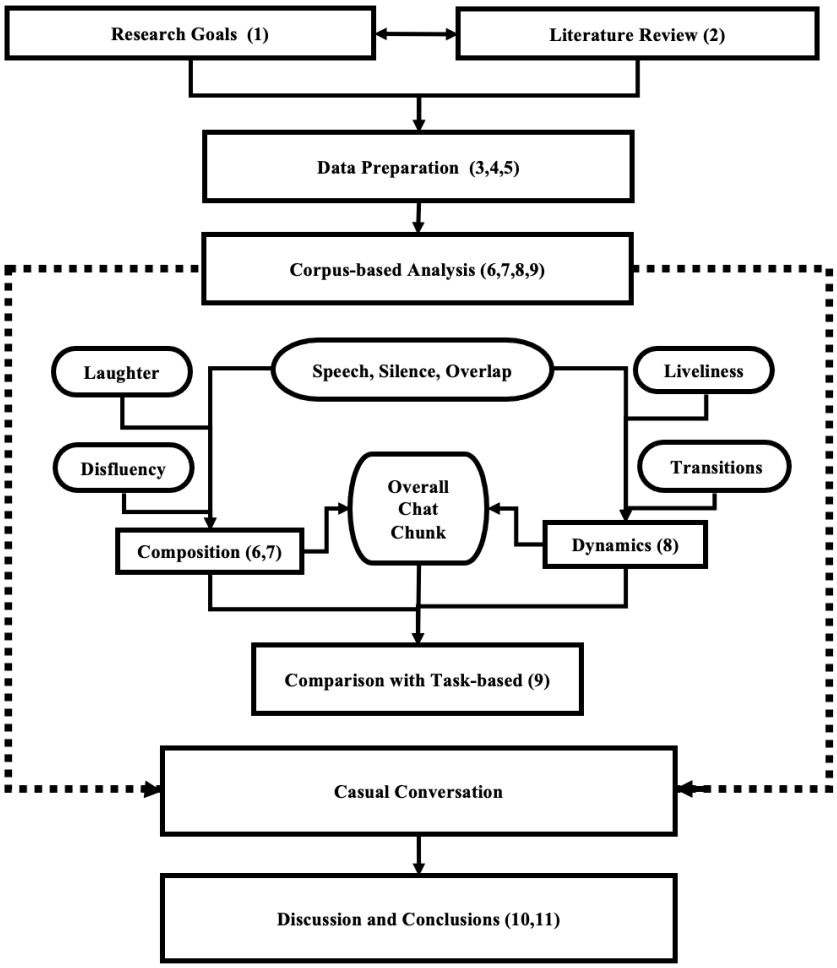


Figure 1.1 – Map of the work covered in this thesis. Numbers indicate the chapters in which the analysis is carried out.

What is the compostition of multiparty casual talk, and of its constituent chat and chunk phases in terms of measures of speech, silence, laughter and disfluency? (1)

What are the characteristics of casual talk, and its constituent chat and chunk phases, in terms of floor state dynamics at the utterance level? (2)

How do casual talk and chat and chunk phases compare to task-based dialogue in terms of speech, silence, overlap, liveliness, and conditions around floor state changes? (3)

These questions are tackled by a series of grouped analyses, addressing the different features of interest. Below is a brief overview of the work performed.

A corpus-based analysis of 8 hours of multiparty data requires that the data be first segmented into speech, silence, and laughter, and then transcribed. The chat and chunk phases described in the literature must be identified in the data for analysis to take place - the corpus must be segmented into chat and chunk phases following definitions in existing literature. This task entails the creation, implementation, and checking of a novel chat and chunk annotation scheme, as described in Chapter 4, Section 4.7, and summarised and discussed in Chapter 10, Section 10.1.

Once the data have been prepared, the work moves on to corpus-based analysis of casual conversations, to get a quantitative picture of the composition and dynamics of casual talk to add to the many qualitative descriptions in the literature. The composition of casual talk is investigated in terms of its basic components - speech, silence, laughter and overlap, as they are distributed overall and in chat and chunk phases, and in terms of how chat and chunk phases are arranged in conversation.

The basic composition of casual conversation in terms of speech, silence, laughter and overlap is analysed in Chapter 6, Section 6.2, and results are discussed in Chapter 10, Section 10.2. Chat and chunk phases are then analysed in Chapter 6, Section 6.5, which explores the characteristics of chat and chunk phases in terms of timing,

duration, and distribution. The composition of chat and chunk phases in terms of proportions of speech, silence, laughter and overlap is then contrasted in Chapter 6, Section 6.5 The results are discussed in Chapter 10, Sections 10.2.2. The analysis of the composition of casual talk is then extended with an analysis of the incidence of disfluencies in a subset of the data, exploring the distribution of disfluencies in casual conversation, and in its constituent chat and chunk phases in Chapter 7, Section 7.2, and results are discussed in Chapter 10, Section 10.2.3.

The analysis then moves to dyamics of casual talk, starting at the local utterance or turntaking level, in order to explore the duration of speech, silence and overlap intervals, and the type of overlap encountered, with experiments in Chapter 8, Section 8.2 exploring the characteristics of chat and chunk phases at the utterance or local to turntaking level in terms of duration and type of utterance, silence, overlap and laughter. Results are discussed in Chapter 10, Section 10.3.1. The level of interactivity or 'liveliness' in casual talk and its chat and chunk phases is also analysed in Chapter 8, Section 8.4, in terms of overlap relative to speech in the clear (compression), and of rates of change of speech distribution across time. Results are discussed in Chapter 10, Section 10.3.2. Within and between speaker transitions, the evolution of the conversational floor state from a stretch of single party speech in the clear to the next stretch of single party speech in the clear by the original or a different speaker, are then investigated and analysed, first across whole conversations, and then at the level of individual chat and chunk phases, in Chapter 8, Section 8.5. Results are discussed in Chapter 10, Section 10.3.3.

The composition and dynamics of casual conversation are then contrasted with a corpus of multiparty task based dialog in Chapter 9, in order to compare casual talk in general, chat and chunk phases, and task-based dialogue in terms of speech, silence, overlap, liveliness, and conditions around floor state changes. Results are discussed in Chapter 10, Section 10.4.

The results of these investigations will provide further insight into the mechanics of multiparty casual conversation, and into the distribution of speakers' contributions in conversations longer than a few minutes. In chat phases of longer conversations, as in very short conversations, contributions should be divided among several participants, while chunks should be common in longer conversations, with individual participants monopolising the conversation while they tell a story or give an extended opinion. The chat segments could be likened to a 'ping-pong' session where participants fire short volleys at one another, while the chunk segments resemble a situation where single speakers hold the conversational 'ball' for extend periods. The questions above aim to elucidate the characteristics of these two phases of conversation, to extend understanding of the nature of casual conversation.

In addition to addressing the questions above, research outputs include:

1. A manually segmented, transcribed, and annotated corpus of 6 long conversations, labelled for speech, silence, overlap, laughter, and chat and chunk phases.
2. An annotation scheme for chat and chunk phases in casual conversation.
3. A novel labelling scheme and analysis paradigm for floor state transitions in conversation.
4. Disfluency labels for one conversation.
5. Evidence in support of or opposition to current theory for casual conversation.
6. Insights to guide design of dialog technology to artificially produce casual talk.

These research outputs will increase the data available to the community; the scarcity of annotated casual conversation data is a major impediment to the advancement of knowledge of how multiparty casual conversation works, and the addition of several hours of segmented, transcribed, and annotated data will be of use to researchers. The chat and chunk annotations and the novel annotation scheme, met-

rics, and analysis methods for floor state transitions will aid in a more detailed exploration of the structure of casual talk and exploration of the realities of turn change and retention in conversation. The disfluency annotations for one of the conversations, very time consuming to perform, are, to the author's knowledge, the only such resource available for multiparty casual talk. In addition all data in CasualTalk were video recorded, so the annotations produced for this thesis can be combined with analysis of phenomena in the video stream for future work on multimodality in casual talk.

The work in this thesis, and the results on the research questions above, should also help to provide evidence for or against existing thinking on multiparty casual talk, much of which is based on 'analysis by exemplar' or on very small scale corpus studies without benefit of precise timing information. Finally, the knowledge gained in the analyses will be useful in designing and building spoken dialog technology which truly provides conversational interaction with human users.

1.4 Thesis outline

This thesis is divided into four parts. Part 1 introduces the research questions addressed and provides background on casual conversation. Part 2 describes the data, annotations, and statistical methods used in the analysis of casual conversation in the thesis. Part 3 describes the experiments comprising the analysis, while Part 4 summarises and discusses the results obtained, and points to future work.

The chapters are briefly described below:

Part 1

- Chapter 2 provides a review of the relevant literature, addressing theories and models of dialogue and conversation, and the various mechanisms which allow spoken interaction to be performed efficiently and seamlessly. The literature on social talk is reviewed with particular focus on work describing casual talk and

distinguishing it from more formal speech genres and written language.

Part 2

- Chapter 3 outlines the challenges of collecting multiparty casual conversational data, and describes the recordings used in the experiments.
- Chapter 4 discusses issues related to the segmentation and annotation of multiparty speech data, outlines how data were segmented, and describes the transcription and annotation schemes implemented and the transcription and annotation of laughter, phase, and disfluency performed on the data.
- Chapter 5 contains an overview of the statistical methods and measures used in the analysis.

Part 3

- Chapter 6 focusses on the composition of talk, presents descriptive statistics for the CasualTalk dataset in terms of speech, silence, overlap and laughter distribution, and contrasts chat and chunk phases in terms of these features.
- Chapter 7 describes a study of the distribution of disfluency in chat and chunk phases of one of the conversations in CasualTalk.
- Chapter 8 focusses on the dynamics of talk, and presents analyses of changes in *floor state*, defined as the totality of who is speaking, silent or laughing at any stage in a conversation, in the CasualTalk datset as a whole. Floor state changes are then contrasted in chat and chunk phases.
- Chapter 9 contrasts the results obtained in earlier chapters with the distribution of state changes in a corpus of task-based conversations.

Part 4

- Chapter 10 summarises and discusses the results of experiments in the preceding chapters.
- Chapter 11 presents the conclusions drawn from the work in the thesis, and outlines future work.

In the interests of reproducible research, all annotations and scripts used in this work are available from `https://github.com/egilmartin/casualtalk.git`

Chapter 2

Related Work

2.1 Introduction

This thesis examines the structure of multiparty casual conversation. In this chapter, theories and descriptions of spoken interaction are briefly described, casual conversation is defined and a review of work in this area is provided. Previous work in corpus studies of social or casual talk is discussed, and features and factors important in analysing casual talk are described, in order to better understand the questions this thesis addresses: how casual talk, with its constituent chat and chunk phases, is organised in terms of speech, silence, conditions around gaps and overlaps, disfluency and laughter.

2.1.1 Conversation - Definitions

Conversation or talk is a universal feature of human society, as indeed is language. Although ubiquitous, human spoken interaction has proven difficult to define exactly; indeed communications sciences in general have been described as lacking a core theory (Berger, 1991). Theories and descriptions of facets of human spoken interaction or conversation have appeared in several disciplines, including linguistics, sociology, anthropology, psychology, information science, and philosophy. This frag-

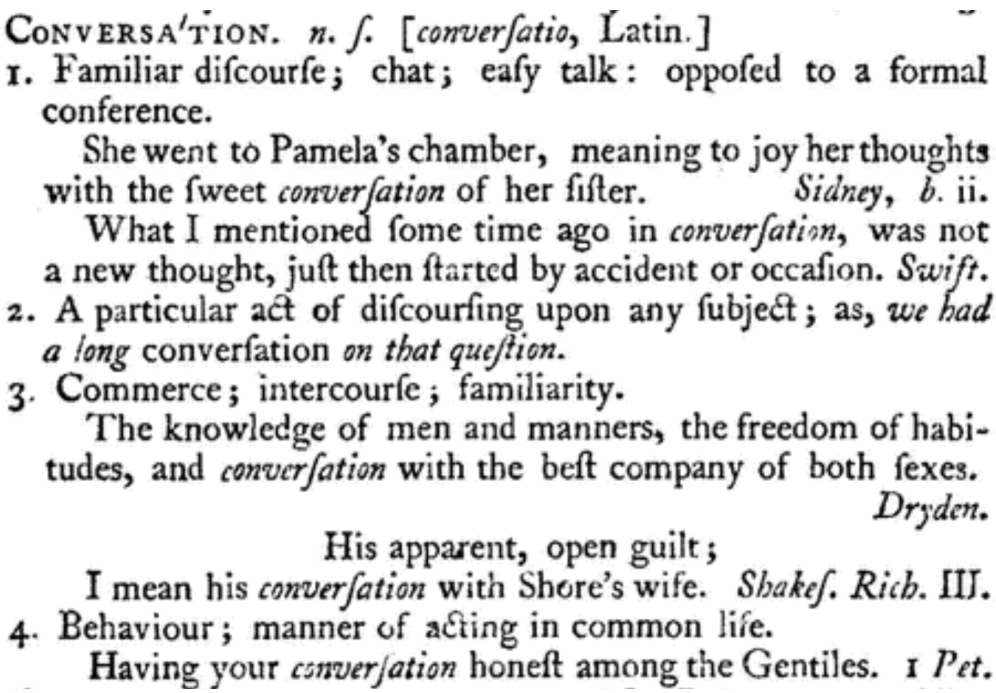
mentation of enquiry into different disciplines has been mentioned as a barrier to a full scientific understanding of conversation (Ginzburg, 2012).

The word *conversation* is derived from 'Middle English (in the sense 'living among, familiarity, intimacy'): via Old French from the Latin conversatio(n-), from the verb conversari' (OED, 2020). By the mid-1700s, the word had taken on its modern sense related to talk exchanged between people. Johnson's 1755 definition of conversation Johnson (1755), shown in Figure 2.1, has the first two entries as 'familiar discourse; chat; easy talk: opposed to a formal conference', and ' a particular act of discoursing upon any subject'. These are followed by definitions related to 'commerce, intercourse, familiarity' and 'behaviour, manner of acting in common life'. The current Lexico (formerly Oxford Living Dictionaries) website defines conversation as a noun meaning 'a talk, especially an informal one, between two or more people, in which news and ideas are exchanged'. [1]

While the word 'conversation' seems to be stably defined by non-technical dictionaries as as largely informal spoken interaction between copresent participants, it is not as clearly defined in many of the disciplines related to the study of interpersonal interaction and communication, where research into 'conversation' can cover a multitude of genres or types of interaction. These have included telephone calls to emergency services, where the participants are not copresent and the purpose is strictly defined, classroom dialogue, which is formal and does not entail equal participation, and televised political interviews, which are often closer to performances before an audience than to spontaneous talk.

As human spoken communication is situated, occurring for different purposes and in different contexts, it is logical to suspect that its characteristics vary with the type of interaction or 'speech-exchange system' (Sacks et al., 1974) participants are engaged in. While types or genres of written communication are fairly clear, categorization is not as straightforward for speech genres. The range of interactions in-

[1] lexico.com/definition/conversation accessed on 18 May 2020



CONVERSA'TION. *n. ſ.* [*converſatio*, Latin.]

1. Familiar diſcourſe; chat; eaſy talk: oppoſed to a formal conference.

She went to Pamela's chamber, meaning to joy her thoughts with the ſweet *converſation* of her ſiſter. *Sidney*, *b.* ii.

What I mentioned ſome time ago in *converſation*, was not a new thought, juſt then ſtarted by accident or occaſion. *Swift.*

2. A particular act of diſcourſing upon any ſubject; as, *we had a long* converſation *on that queſtion.*

3. Commerce; intercourſe; familiarity.

The knowledge of men and manners, the freedom of habitudes, and *converſation* with the beſt company of both ſexes. *Dryden.*

His apparent, open guilt;
I mean his *converſation* with Shore's wife. *Shakeſ. Rich.* III.

4. Behaviour; manner of acting in common life.

Having your *converſation* honeſt among the Gentiles. 1 *Pet.*

Figure 2.1 – Definition of 'conversation' from Johnson's 1755 Dictionary of the English Language

volving speech which humans engage in is enormous, with the problem of categorizing different types of speech exchange into genres described as 'notorious' (Bakhtin, 1986). Types of spoken interaction or 'speech-exchange systems' range from highly formal rituals, through lectures or classes, business negotiations, interviews and question and answer sessions, to casual talk between friends and family or strangers. A major goal of research into spoken interaction is to observe and analyse the dynamics of changes in the floor state – who is speaking at any point, and the distribution of speech and silence among participants.

Investigation of these conversational dynamics is often based on corpora comprising collections of (transcribed and annotated) recorded speech. Many of these are recordings of task-based interactions, either real or staged as role plays, while there is a dearth of recordings of naturally occurring casual conversation. Task-based activities include business meetings (both real (Janin et al., 2003) and staged (Mc-

Cowan et al., 2005)), information gap games such 'spot the difference' (Baker and Hazan, 2011) and 'balloon tasks' where participants prioritize items on a list (Vinciarelli et al., 2012), or real-life emergency line and suicide hotline recordings such as those collected by pioneers such as Harvey Sacks , Emmanuel Schelgoff, and Gail Jefferson (Schegloff and Sacks, 1973; Sacks et al., 1974). Although such corpora have been used to analyse 'conversation', it is not certain that one size fits all for interactional data; that results and insights gained, for example, from measurements of the timing of pauses and gaps in telephone conversations about a 'balloon task' or indeed in real-life suicide hotline recordings, are generalizable to other genres (Lemke, 2012), such as casual conversation, where the rules of participation may differ. This question is further discussed and an overview of existing data is given in Chapter 3.

For the purposes of this work, casual conversation is defined as informal face-to-face conversation between co-present participants, planned and produced spontaneously, where the purpose of the talk is not to perform a practical task. Below, the origins of face-to-face spoken interaction and conversation are briefly explored.

2.1.2 Face-to-Face Interaction

For most of human history, communication has been person to person, using voice and gesture, with different communication methods and tools appearing quite recently. Human communication has been broadly divided into two systems, each transmitting information of differing types. The first, and older, is the gesture-call system (Burling, 2007), a vocal and visual communication system common to primates, which includes facial expressions, cries, and some physical gestures. This system has been studied by anthropologists and social psychologists (Dunbar, 2003; Mehu et al., 2007), and has been understood to transmit predominantly affective or emotional information, but has also been argued to communicate propositional content in the form of warnings (Seyfarth et al., 1980). The system requires co-presence – with par-

ticipants in close proximity for expression or gesture, or within earshot for vocal calls. The second system is language, a more recent phenomenon, commonly estimated as at least 100,000 years old (Nichols, 1998; Perreault and Mathew, 2012) although genetic findings on the possibility that the FOXP2 or 'language' gene was present both in early humans and in Neanderthals may push the date for the emergence of language back to around 1.8 million years ago (Krause et al., 2007). Both systems are active in face-to-face conversation.

Vocal expression and interaction were present in primates long before human language evolved, and vocal interaction involving both language and gesture-calls was present in humans for tens of millennia before the relatively recent invention, c. 3500 BCE, of writing, a code which allows the transmission of linguistic information via media other than the voice, and thus makes possible the preservation of such information and its diffusion across time or space. Several researchers have stated a belief that interpersonal spoken conversation is the primary or unmarked case of human communication using language, which differs from many other uses of language as it is planned and produced spontaneously (Dunbar, 1998a; Abercrombie, 1956; Hayakawa, 1990; Turnbull, 2003), often mentioning its fundamental role in society. An example of these views is Turnbull's definition of conversation in the context of Social Pragmatics:

> 'Face-to-face talk is a universal form, and probably the basic form, of social interaction. Other forms of talk, such as telephone talk and therapy talk, are variants of this basic form of social interaction. Face to face talk between two or more persons is the primary site of sociality in all cultures. It is the place where persons together negotiate meaning and personal, social and cultural identities. Talk is situated action: it is action occurring manifestly and locally between two or more specific persons. Further talk cannot be achieved without the mutual assumption that each participant

> will respond to a prior turn with a conditionally relevant turn. Thus talk is not possible without trust. Overall, then, social pragmatists view talk as a manifest, co-constructed, for-all-practical purposes, indexical, moral, achievement' (Turnbull, 2003, 179)

Face to face spoken interaction comprises a bundle of visual and audio signals and cues, transferring propositional content, affective and attitudinal information through words, physical and vocal gesture, silence and laughter. However, a large part of the background research into conversation has stemmed from disciplines which treat spoken interaction as oral transmission of linguistic text and concentrate on the linguistic context of utterances, without regard to other factors.

Below is a brief overview of contributions from the various disciplines and sub disciplines where attempts to describe conversation have been made.

2.1.3 Approaches to Spoken Interaction

Although spoken interaction is the most basic use of linguistic communication by humans, it remained surprisingly under-researched until the final third of the 20th century. As a subject of study it has not received as much interest as more formal written language throughout the history of communication studies and linguistics. As writing evolved, it became the academic medium for discussion and transmission of ideas, and the study of language itself has been based on the written word for much of its history - indeed the word grammar is derived from the Greek $\gamma\rho\alpha\mu\mu\alpha$, 'that which is written, a written character, letter' (Liddell et al., 1966). Many researchers have highlighted 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). While there has been attention paid to the *sounds* of language in phonetics and phonology research for some time, the structure and content of spoken conversation was

generally ignored (Halliday, 1989). Indeed, it has been pointed out that phonetics and phonology research should use casual conversation data to reflect the realities of spoken language, and that work on mental processes in speech should also focus on natural conversation rather than isolated model phrases in order to better investigate how humans process language (Tucker and Ernestus, 2016).

According to Chafe, much of the lack of comprehensive theory on human spoken interaction has been attributed to a 'literary' bias in academic consideration of language and spoken interaction (Chafe, 1994). In this 'literary' view, speech is seen as a medium for the transmission of language, and language is seen as a collection of well formed phrases or sentences, most faithfully expressed via the written word, a view which underpins Saussure's notions of langue and parole or Chomsky's competence and performance paradigm (Saussure, 1916; Chomsky, 1965). This overturning of the primacy of oral communication, and in particular the phrase 'oral literature', led Ong to remark that the treatment of 'oral literature' as a variant of writing was akin to 'thinking of horses as automobiles without wheels' (Ong, 1982). Gleick stresses how difficult it is for a literate society to contemplate preliterate communication. He also mentions that literacy has greatly extended the vocabulary of language - with English containing over a million words while the average oral language has a vocabulary of a few thousand words (Gleick, 2011). This could indicate that written language has led the expansion of the propositional information transfer function of language, and indeed of human thought (Ong, 1982). This adaptation towards the use of written language as the medium for transmission of ideas could also be a factor in the literary bias, with scholars not realising the role that 'para-' and 'extra-' linguistic factors (such as prosody, gesture, and context) play in face-to-face spoken interaction, as Hayakawa rather scathingly states in his description of how reliance on text-based enquiry fails to address spoken language:

'people who have spent much of their lives in the study of written symbols

> tend to be relatively deaf to everything but the surface sense of words'
> (Hayakawa, 1990, 77)

Another factor in the lack of study of spoken interaction compared to written is the ephemeral quality of spoken interaction – while written text is a concrete record in itself, amenable to extended study, conversations and indeed all forms of speech do not leave an authoritative source trace unless they are specifically recorded.

In discussion of written and spoken language, it is important to note that there is no single phenomenon 'writing' or 'speaking', but rather that there are several different types of written and spoken language, or more accurately, several different settings or activities which are mediated in different ways by language (Chafe and Danielewicz, 1987) in a space akin to Finnegan's 'oral-written continuum' (Tannen, 1982, 1985). In terms of use, Olson's notion that the meaning of written language is in the text while that of spoken language is in the context seems to be uncontroversial, when we consider informal conversation versus expository text (Olson, 1977). However the lines have been becoming rather blurred over the last few decades. Print media intended for immediate consumption such as advertising and the popular press have been observed to use a style close to conversational language, a process termed 'conversationalization' (Fairclough, 1992). With the advent of computer mediated communications such as email, bulletin boards, and digital social media there has been a further upsurge of casual writing - writing which is not performed for a formal purpose but rather to fulfil social goals. This writing uses structures and content found in casual speech, using emoticons and exaggerated punctuation to convey prosody and affect, and intermingling text with images and video. Unlike traditional means of communication in writing, it is synchronous, and thus similar to speech, as it is instantly published and open to immediate reply. However it is recorded and therefore lacks the transience of speech - a feature of conversation noted by Halliday's observation that

'Writing exists whereas speech happens.' (Halliday, 1989, xxxiii)

A useful overview of differences between some of the genres of written and spoken language is given in Chafe and Danielewicz's 1987 report to the US National Centre for the Study of Writing (Chafe and Danielewicz, 1987), where they catalogued differences between conversation, academic lecturing, informal letters and academic writing in a corpus of the four genres collected from 20 lecturers and students. They explored lexical choice, clause length and construction, sentence structure, and inherent social interaction. For conversation, they found relatively limited vocabulary with extensive hedging and use of colloquial language, forming brief intonation units, which were chained to form longer stretches of speech, delivered in a highly interactive manner. They concluded that such talk was highly constrained by the processing requirements of listeners. For academic lectures, they found similar restricted lexical choice but higher levels of literary vocabulary, with only slightly longer intonation units than conversation, but less interaction with the audience. They concluded that lecturing was a mixed form of language - constrained by listeners' processing faculties and thus retaining the chaining of short clauses present in conversation, but 'striving after some of the elegance and detachment' of formal language. (Chafe and Danielewicz, 1987, 26). They found letter writing vocabulary more varied, hedging was rare in this genre, there was a moderate use of colloquial language and more literary language appeared than in the spoken data. They also reported the highest level of self-involvement in this genre. In academic writing, they found maximally varied vocabulary, avoidance of hedging and inexplicit references, and almost no colloquial items or contractions. Intonation units were 'maximally' long and used all literary devices which could expand sentences while maintaining coherence.

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 40 million word Longman Corpus of Spoken and Written English (LSWE), and

particularly in their chapter on the grammar of conversation(Biber et al., 1999, 1037-1125), where they outline differences in the grammar of spoken and written English based on a large-scale study of the corpus. This work is further described in Section 2.3.3 below.

Work of the type described in this section has elucidated some of the characteristics of casual conversation by differentiating it from other genres in terms of linguistic content and characteristics of utterances. The work in this thesis aims to add quantitative information on the timing of multiparty casual conversation to this body of knowledge. A brief description of models of dialog and conversation is given below.

2.2 Models of Dialog and Conversation

Several models of spoken dialog and conversation have been proposed - ranging from 'encoder-decoder' analogies from communications theory to models of conversation as a product of joint activity where interlocutors co-ordinate mental representations by updating 'common ground'. Important models are briefly sketched below.

2.2.1 Code Messaging vs. Speech Acts

An influential perspective on spoken interaction was provided by the 'code' model of talk, from Shannon's General Communication Model (Shannon and Weaver, 1948). In this model, a sender codes a message to form a signal which is then transmitted across some (possibly noisy) channel to a receiver who decodes the message and then formulates a message in response. This model fits with the Saussurean speech chain notion - of a speaker performing sentences as utterances which are picked up by the listener and decoded using linguistic competence (Saussure, 1916).

A fundamental question is whether the information encoded in spoken language reflects representations of the world or rather the intentions of the speaker and how they intend the recipient to act. We appear to *do* things with speech. This observa-

tion is the basis of Speech Act Theory (Austin and Urmson, 1978; Searle, 1985), where utterances are analyzed in terms of what they cause to happen in the world, as performatives, considered in terms of what we are speaking *for* rather than what we are speaking *about*. The notion of meaning was further questioned by Grice's notions of natural and non-natural meaning, where natural meaning is the 'literal' meaning of an utterance and non-natural meaning refers to the implicature or inferred meaning of an utterance (Grice, 1975).

2.2.2 Dialogue as a site for co-construction of meaning

In contrast to the code model, many researchers theorise that spoken interaction is a joint activity, where meaning is co-created incrementally between the interlocutors, rather than fully formulated by the sender, encoded, sent, and decoded by the receiver. In this view, conversation is seen as a series of sequences where participants collaborate to build meaning. Participants need to design their utterances to reflect the recipient's affordances (how much to transmit per utterance, how to present the information semantically, syntactically, prosodically, and pragmatically) to allow grounding of the information transmitted. Participants constantly update the information state not only with what has been said but also with whether it has been received and understood as intended - this interaction relies on a system of backchannels and nods from the listeners; and prosodic devices by the speaker, including pauses and intonational patterns, to ensure that content is being understood.This incremental process of grounding has formed the basis of much work on dialogue (Clark, 1996; Brennan and Hulteen, 1995).

2.2.3 Ethnomethodology and Conversation Analysis

Much work on spoken interaction or 'talk in interaction' arose from sociology, particularly the work of Garfinkel, Goffman and Sacks, Shegloff, and Jefferson, who ex-

amined spoken interaction or conversation as an example of social interaction. Goffman viewed social interactions as social institutions with a particular order or syntax - as an arena in which participants manage their projection of face, self, and identify. Here, the interaction itself is the focus of attention, rather than the linguistic content of the utterances involved. In this view, according to Kendon, an 'occasion of interaction such as a conversation' could be seen as an occasion where 'participants enter into a complex system of relationships' which are 'discoverable and generally applicable' (Kendon, 1990, 4).

Garfinkel, reacting to the Parsonian view of co-ordinated action as the product of compliance with external social norms, asserted that social actors built their common view of the social world through interaction and shared practical reasoning or 'ethnomethods', and that the social context and actions affected each other to build a social reality. Ethnomethodology was based on observation and description of real interactions, of which talk or conversation was a prime example. Sacks, Shelgoff and Jefferson initially studied telephone calls to the emergency services as examples of orderly interaction, later expanding their work to investigate the mechanics and processes of spoken interaction more generally, forming the discipline of Conversation Analysis.

Conversation Analysis describes conversation as a series of utterances or talkspurts arranged in sequences of turns by different parties (Sacks et al., 1974; Clark, 1996; Schegloff, 2007). This turntaking has been ascribed to the fact that it is difficult for humans to separate verbal content in voices speaking at the same time, and it is seen as a basic component of all social organization (Sacks et al., 1974). In addition to this temporal organization, described as a mere 'acoustic fact' or artefact of engineering contingencies, there is an independent interactional ordering into sequences composed of adjacency pairs, a 'social fact' (Schegloff and Sacks, 1973). Adjacency pairs consist of two part sequences where the first part limits the range of

possible second parts, thus providing predictability and organization to conversation. The simplest example is a question, which narrows expectation of the form of the next contribution to an answer. Other first and second part pairings include greeting-greeting, complaint-denial, and assertion-acknowledgement. Adjacency pairs can be nested within other adjacency pairs particularly when misunderstandings give rise to repair sequences.

Human speech is composed of words, created from the phonemes of the language, and also of a range of other tokens including interjections, inserts, backchannels, filled pauses, hesitations, and disfluencies. The form and function of words, phrases, and these tokens vary in different speech situations. At any moment, each participant in a conversation may be speaking, making non-speech sounds, or silent, giving rise to a large number of possible conversational floor states. Conversation is peppered with silences of different types - gaps, pauses, and lapses. Speakers may overlap other speaker's speech during or at the end of a turn as feedback, or competitively at turn relevant places (TRPs). The form and distribution of all of these phenomena may help to classify different speech genres, and analysis of conversational floor state in multiparty casual talk forms a major part of the work in this thesis.

It is clear that notions of spoken conversation have moved from a purely code model to a collaborative view, where interactants participate in joint activity. A pertinent question is what does spoken interaction accomplish? Below the notions of practical or instrumental and interactional (interpersonal) or social talk are described.

2.3 Interactional and Instrumental talk

As mentioned above, human spoken communication is situated, occurring for different purposes and in different contexts, and its characteristics vary with the type of interaction or 'speech-exchange system' (Sacks et al., 1974) participants are engaged in.

Many studies of spoken interaction suggest that spoken interaction can be broadly divided into two categories - 'instrumental' or task-based dialogue where the focus is on information transfer, and 'interactional' or interpersonal, casual, or social talk, which can manifest as a number of genres, including 'small-talk' or chat, gossip, discussion, 'dinner-party' conversation, and conversational narrative.

Much attention has been paid to 'task-based', 'instrumental' or 'transactional' dialogues such as buying a pizza, asking for information, or participating in a job interview, rather than the 'unmarked' case of social talk (Schneider, 1988). In task-based interactions there is a clear goal or 'point' to the interaction, which can be completed within the current conversation. Participants are aware of this goal, and success is often marked by reaching the goal efficiently, although success may also be partially driven by factors such as the 'pleasantness' of the encounter.

Much real conversation is not clearly task-based, and may be better described as interactional. In these conversations, there is often no obvious short term task to be accomplished through speech and the purpose of the interaction is better described as a shorter or longer term goal of building and maintaining social bonds (Malinowski, 1923; Dunbar, 1998b). In such interaction, henceforth *casual conversation*, the transfer of verbal information is not always the main goal, and indeed the simple fact of speaking, Malinowski's 'mere exchange of words' (Malinowski, 1923), may carry greater importance in attaining the goals of social bonding.

As an example, a tenant's short chat about the weather with the concierge of an apartment block is not intended to transfer important meteorological data but may help to build and maintain a relationship which may serve either of the participants in the future. Similarly, conversation exchanged by a group of friends meeting for dinner serves to maintain and further define their relationship rather than meet practical goals - while it may be true that these bonds serve the participants in the relationship in practical ways, talk occurring during congregation of people seems to be a natural

activity arising without strategic planning on the part of participants. This notion of conversation as a marker of the status of interpersonal relationships appears in common language; we say 'They're not speaking' of people who have a relationship, to mean that their good relations have broken down, while 'They're not speaking about X' does not carry the same information on interpersonal relationships, but rather describes an aspect of the propositional content of a dialogue.

Of course, individual conversations can include elements of both instrumental and interactional talk- indeed much business talk contains stretches of 'small talk' or chat, which is believed to 'oil the wheels' and build good business relations.

The study of different spoken interaction types or 'speech exchange systems' can help discriminate which phenomena in one type of speech exchange system are generalisable and which are situation or genre dependent. In addition to extending our understanding of human communication, such knowledge will be useful in human machine interaction design, particularly in the field of 'companion' robots or relational agents, as the very notion of a companion application entails understanding of social spoken interaction. Indeed, spoken dialogue system technology is based on task-based dialogue (Allen et al., 2001a) for reasons of tractability, as early researchers recognised that open domain talk without a clear short term goal was not as amenable to implementation as shorter and more predictable transactional exchanges.

This thesis examines multiparty casual conversation, where there is no previously assigned topic, or task to be accomplished, other than to 'pass the time'. The next section provides a brief overview of existing work in this area.

2.3.1 Approaches to Analysis of Casual Conversation

Casual conversation has been described as 'talk just for the sake of talking' (Eggins and Slade, 2004, 6). Researchers in several fields have mentioned casual conversation as a genre or type of speech exchange system. The descriptions have included

accounts of 'phatic' communion and communication, small talk, chat, and casual or informal conversation.

Malinowski, working in the ethnographic/anthropological tradition, drew attention to 'phatic communion', which he defined as an activity comprising free aimless social conversation (Malinowski, 1923). He proposed that the purpose of this kind of spoken interaction was not to exchange practical information or to express thought, but that this phatic communion was rather an emergent activity of congregating people, which allowed social bonding by first avoiding an uncomfortable silence which could be seen as unfriendly, and then engaging participants in a reciprocal flow of words - generally positive platitudes, comments on the obvious, and personal views and life history. This function Malinowski saw as the primitive or most basic use of speech, existing in parallel to other more sophisticated uses of language as a repository and controller of ideas.

Laver, in his review of the functions of phatic communication (Laver, 1975), bemoans the fact that Malinowski's naming of the phenomenon seemed to have given researchers the impression that no further investigation was necessary, while speculating that the prominence of competence based or mentalistic approaches to language in the 1960's and 1970's, and the resulting lack of attention to communication in action, had discouraged much research. He also concludes that so-called phatic or empty talk is not completely empty of meaning but rather communicates social information allowing interlocutors to relate to one another in terms of their social standing. He thus adds a third social information transfer function to Malinowski's functions of defusing silence and maintenance of positive social bonds. Interestingly, Senft's later criticism of some of Malinowski's claims as to the lack of meaning in the greetings formulae of Trobriand islanders would add strength to Laver's view, as Senft ascribes more socially relevant ingroup information transfer to some elements of what had been described as 'phatic communion', adding strength to Laver's claim

that such talk does in fact have an information transfer form (Senft, 2009).

Researchers in fields including anthropology, evolutionary psychology, and communication have theorized that such talk functions to build social bonds and avoid unfriendly or threatening silence, rather than simply to exchange information or express thought. 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, 1998a). 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). Jakobson, in building his theory of communication (Jakobson, 1960), defined phatic communication (as opposed to communion) as language which maintains the communication channel, as a communication management function, described as corresponding "to that function of language which is channel-oriented in that it contributes to the establishment and maintenance of communicative contact" (Lyons, 1977, 34). Brown and Yule broadly divide dialogue into transactions and interactions (Brown and Yule, 1983).

Transactions are task-oriented exchanges, such as buying a paper. However, even seemingly task-oriented interaction is often replete with interactional talk, with McCarthy reporting that only 10% of the talk in hairdressers was transactional and thus recommending that researchers seriously consider the interactional elements of transactional talk as they probably contribute greatly to transactional success (McCarthy, 2014). Hudak and Maynard observed social talk in surgeon-patient interactions, defining such talk as 'referentially independent from their institutional identities as patients or servants, oriented instead to an aspect of ..personal biography or to some neutral topic ... (e.g.weather), and reporting that such talk aided in the medical task at hand by allowing patients to 'disattend' to potentially embarassing or invasive med-

ical procedures' (Maynard and Hudak, 2008). Researchers have noted that hard and fast categorization of small talk and instrumental talk is not always possible, and that interactional talk permeates many ostensibly practical or instrumental encounters (Benwell and McCreaddie, 2016).

While there has been corpus based study of task-based dialogue and of stretches of small talk within and around core or instrumental talk, there are few such studies of interactions consisting entirely of casual or social talk. There are several possible reasons for this – a major factor is scarcity of data, as will be discussed in Chapter 3, another factor is that casual conversation has often been seen as inconsequential chitchat or 'off-talk' which decorates but does not contribute to the main purpose of the spoken interaction, indeed, subjects asked to record talk at a workplace tended to believe that researchers wanted 'real' talk, and often turned on the recorder only after small talk had ended and the 'real' meeting was beginning (Holmes, 2014). In addition, until very recently, dialog technology was not oriented towards interpersonal talk but rather concentrated on task-based interaction. Below, work concentrating on casual talk as a speech-exchange system in itself is reviewed.

2.3.2 Corpus Studies of Casual Conversation

Schneider performed a study of small talk in the 1980's, based on a corpus of audio recordings of 56 small talk conversations made 'in the wild' throughout Britain (Schneider, 1988). He concentrated on the linguistic content and structure of the dialogues, using discourse analysis principles to describe instances of smalltalk at several levels, from frames such as 'WEATHER' to sequences and adjacency pairs within interaction, and then to the types of utterances comprising these sequences. From these descriptions he built an overall model of how smalltalk progresses and its governing principles. Schneider contrasts 'offensive' and 'defensive' small talk, where he takes offensive to mean talk actively seeking to build relationships (chatting up a

stranger at a party) while defensive refers to talk which avoids embarrassing silence but does not seek to build closer ties (such as conversation between strangers in a waiting room). He incorporates these notions into his 'politesse' and 'friendliness' maxims of small talk. Schneider notes that although Grice's Co-operative Principle may hold for small talk, the Gricean maxims of quantity, quality, relation, and manner do not seem to be well suited as these maxims are primarily tied to the quality and efficient transfer of information. He proposes a politeness principle ('Be polite!') and two super-maxims of politesse and friendliness governing defensive and offensive small talk. Politesse is essentially an avoidance strategy with maxims including Speech ('Avoid silence!'), and Person ('Avoid curiosity!'), while Friendliness is more active including Speech ('Say something nice!), and Person ('Show interest!) (Schneider, 1988, pp157-159). In this way politesse is associated with avoiding embarrassing silences in situations where strangers must spend time together while friendliness is connected to social events where relationships are built. Using these principles as guidelines he describes polite and friendly exchanges in terms of topic, opening and continuing utterance types (in syntactic terms) and utterance content (as speech acts). He identifies sequence types widely used in small talk such as 'remark-agree' which are often followed by idling sequences of repetitions of agreeing tails such as 'Yes, of course', 'MmHmm.'. In terms of topic, he distinguishes the 'safe' topics used in polite talk from more personal topics in friendly talk. He highlights the importance of agreeableness to both types of smalltalk, mentioning a tendency to exaggerate agreement and positive evaluations, and how shared experience is often sought in friendly conversation.

The structure of casual talk at the 'macro' level of an entire conversation was formulated by Ventola (Ventola, 1979), who used transcriptions of four conversations ranging from two to eighteen minutes in length she had recorded surreptitiously in the United Kingdom to identify internal elements or phases of such talk beyond the

social niceties of greeting and leavetaking. She described how a non-transactional or social conversation may comprise several structural elements or phases including greetings, casual or light conversation, and more centred discussion on a particular topic. Ventola's proposed structures for phases of casual talk are more fully described in Section 2.3.4

Casual conversation is viewed as the constructor of social reality by Slade and Eggins (Eggins and Slade, 2004), who define casual conversation as the type of talk engaged in when 'talking just for the sake of talking'. However, they strongly deny any notion that casual conversation is somehow light or 'meaningless', arguing rather that it is through casual conversation that people form and refine their social reality. This 'social work' uses language to negotiate social reality and interpersonal relations. In their work Slade and Eggins define casual conversation in contrast to 'pragmatic' or 'instrumental' talk (Eggins and Slade, 2004). They describe two types of casual conversation - solidarity motivated and difference motivated. They provide examples of gossip, where participants reaffirm their solidarity or ingroup status by jointly ascribing outsider status to another - the subject of the gossip, who paradoxically may be a friend to many or all of the participants. Gossip often takes place in settings where participants who may not be close friends have to get on with one another - such as workplaces. Eggins and Slade also show examples of conversation between closer friends at a dinner party where greater intimacy allows greater differences of opinion. They also identify story-telling as a frequent genre in conversation. They highlight the contribution of both 'chat' (interactive exchanges involving short turns by all participants) and 'chunks' (longer uninterrupted contributions) to the content of conversation. They argue that all three situations - dinner table conversation, workplace gossip, and anecdote swopping can be classified as casual conversation, in contrast to 'pragmatic' spoken interaction, as in an exchange at a post office where one participant is a customer buying stamps and the other is the clerk selling the stamps.

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. They note that 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.

Slade and Eggins summarise the differences between the two classes of talk as:

- Pragmatic talk is well defined by its goals while casual talk negotiates social reality.
- Pragmatic interactions tend to be short while casual conversations can go on indefinitely.
- Casual conversation is often between multiple participants all of whom are equal in role although all do not necessarily participate to the same level. Pragmatic interactions and many of the examples used in conversation analysis often involve two participants with clearly defined and often complementary roles.
- Pragmatic talk, although it may contain social parts, is more formal than casual talk.
- Casual talk contains a lot of humour.

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. The experimental work in this thesis is largely based on the structure of casual talk at the level of chat and chunk phases, as defined by Slade and Eggins. Below is a brief description of work on all aspects of structure in casual talk, followed by a more detailed treatment of Slade and Eggins' chat and chunk model and Ventola's phase model of casual conversation.

2.3.3 The Structure of Casual Talk

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 40 million word Longman Corpus of Spoken and Written English (LSWE), and particularly in their chapter on the grammar of conversation (Biber et al., 1999). This work resulted in a descriptive grammar of sectors of the corpus representing four core registers of contemporary (since 1980) American and British English, each consisting of around 5 million words - conversation, fiction, news, and academic prose. The total 40 million plus word corpus also included two supplementary registers: non-conversational speech and general prose. The main conclusion is that conversational spoken language cannot be adequately described using the grammar of the written language, and the sentence in particular, as a basic unit of analysis.

> 'Whereas the sentence has been treated, traditionally and in modern theory, as the fundamental structural unit of grammar, such a unit does not realistically exist in conversational language.'

(Biber et al., 1999, 1039)

Based on the corpus study, Biber proposes a grammar of conversation based on clausal and non-clausal C-units, where clausal units refer to independent clauses and any dependent clauses within them. Non-clausal units are phrases which do not qualify as clauses. These phrases, often referred to as fragments (Schlangen, 2009; Fernández and Ginzburg, 2002), make up over two-thirds of the conversational speech section of the LSWE. As an example of how these units may be produced by a speaker, Thornbury describes the following typical utterance taken from Halliday (Thornbury and Slade, 2006):

> 'but you can't get the whole set done all at once because if you do you won't have any left to use at home, unless you just took the lid in and

kept the boxes, in which case you wouldn't have to have had everything unpacked first; but then you couldn't be sure the designs would match so...'

(Halliday, 1989, xxiv)

Thornbury describes the example as produced 'on the fly' as a series of stages, based on evidence from psychology that human working memory is normally limited to around seven items (whether words or multi-word units) (Miller, 1956), and thus would not allow for planning of this long phrase. He suggests that the utterance may be understood as a successive accumulation of individual clause-like units, rather than embedded clauses as would be more common in writing (which of course allows for editing before sending) (Thornbury and Slade, 2006, 77).

'but you can't get the whole set done all at once

+ because if you do

+ you won't have any left to use at home

+ unless you just took the lid in and kept the boxes

+ in which case you wouldn't have to have had everything unpacked first

+ but then you couldn't be sure the designs would match

+ so...'

This recalls earlier analyses of written versus oral expression (Ong, 1982; Chafe and Danielewicz, 1987) which stressed the cumulative right-additive nature of spoken expression. A more elaborate demonstration is given in Halliday's later work, where he takes several samples of spoken conversation and shows how they would need to be altered or re-engineered to create plausible written text(Halliday, 1994, 54-55). Halliday notes the complexity of spoken utterances comprising multiple lengthy

clauses, describes how they were difficult to follow in writing, but stresses that in speech such utterances were remarkably well formed, and easily understood by a listener:

> 'although the speaker seemed to be running through a maze, he did not get lost, but emerged at the end with all brackets closed and structural promises fulfilled... while the listeners had absorbed these passages quite unconsciously and without effort, they were difficult to follow in writing'

(Halliday, 1994, 55)

2.3.4 Phases of Casual Conversation

In terms of structure, it has been noted that casual conversation can run indefinitely, and thus there is a need for analysis of longer conversations. Conversations have been described in terms of segments longer than adjacency pairs or exchanges. Ventola describes the 'superstructure' of casual conversation 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). Eggins and Slade describe casual conversation as proceeding in 'chat' and 'chunk' phases, where chat is highly interactive and chunk comprises more monologic input where one speakers retains the floor for an extended period, often telling a story or giving an opinion (Eggins and Slade, 2004). These approaches to describing the structure of casual talk are further described below.

Ventola's Phases

The form of casual talk at the 'macro' level of an entire conversation was formulated by Ventola (Ventola, 1979), who described how a non-transactional conversation may

comprise several structural elements or phases. Ventola's phases could well be called generic in Bakhtin's terms, particularly in light of his idea of 'polyphony' or cultural priming of speech-exchange systems:

> 'we learn to cast our speech in generic forms and when hearing others' speech, we guess its genre from the very first words; we predict a certain length and a certain compositional structure; we foresee the end; that is from the very beginning we have a sense of the speech whole, which is only later differentiated during the speech process. If speech genres did not exist and we had not mastered them, if we had to originate them during the speech process and construct each utterance at will for the first time, speech communication would be almost impossible'

(Bakhtin, 1986, 83)

The structural elements she described are:

G Greeting.

Ad Address. Defines addressee ("Hello, Mary", "Excuse me, sir") - rare with strangers but possible.

Id Identification (of self) - only for strangers.

Ap Approach. Basically smalltalk. Can be direct (ApD) – asking about interactants themselves (so usually people who already know one another), or indirect (ApI) – talking about immediate situation (weather, surroundings, so can happen between strangers or with greater social distance). These link back to Laver's psychological treatment of the margins of talk, and forward to Schneider's maxims for offensive and defensive smalltalk. In Ventola's view, these stages allow participants to get enough knowledge about each other to enter more meaningful conversation.

C Centring. Here participants become fully involved in a conversation, talking at length. This stage is much less predictable than the Approach stage in terms of topic, and can range over several overlapping topics for an indeterminate number of repetitions of the stage.

Lt Leave-taking. Signalling desire or need to end conversation.

Gb Goodbye. Can be short or extended, in which case there are projections to further meetings.

Ventola develops a number of sequences of the elements above characteristic of conversations involving different levels of social distance. She also describes conversations as minimal or non-minimal, where a minimal conversation is essentially phatic, particularly in Jakobsen's sense of maintaining channels of communication (Jakobson, 1960) - such a conversation could simply be a greeting, or it could be a sequence of approach stages. Non-minimal conversations involve centring – where the focus shifts to information being exchanged. Ventola stresses that this distinction should not be temporal, but is structural and based on the organisation of information content in the talk.

Several of the elements are optional and omitted in particular situations. For example, friends can jump from Greeting to Centring without passing through the 'smalltalk' exploratory Approach stages when having a longer conversation. Strangers would not normally greet one another but could start with ApI ("It's a nice day").

Some elements usually appear only once (G (greeting), Gb (goodbye)) but others can recur. Approach stages can occur recursively generating long chats without getting any deeper into centring. Centre stages can recur and indeed can be interspersed with approach stages. Figure 2.2 shows a simplified example of a possible chain in Ventola's scheme drawn as a directed graph.

Ventola has several possible chains of elements to represent different conversation types between friends and strangers. The symbols used are:

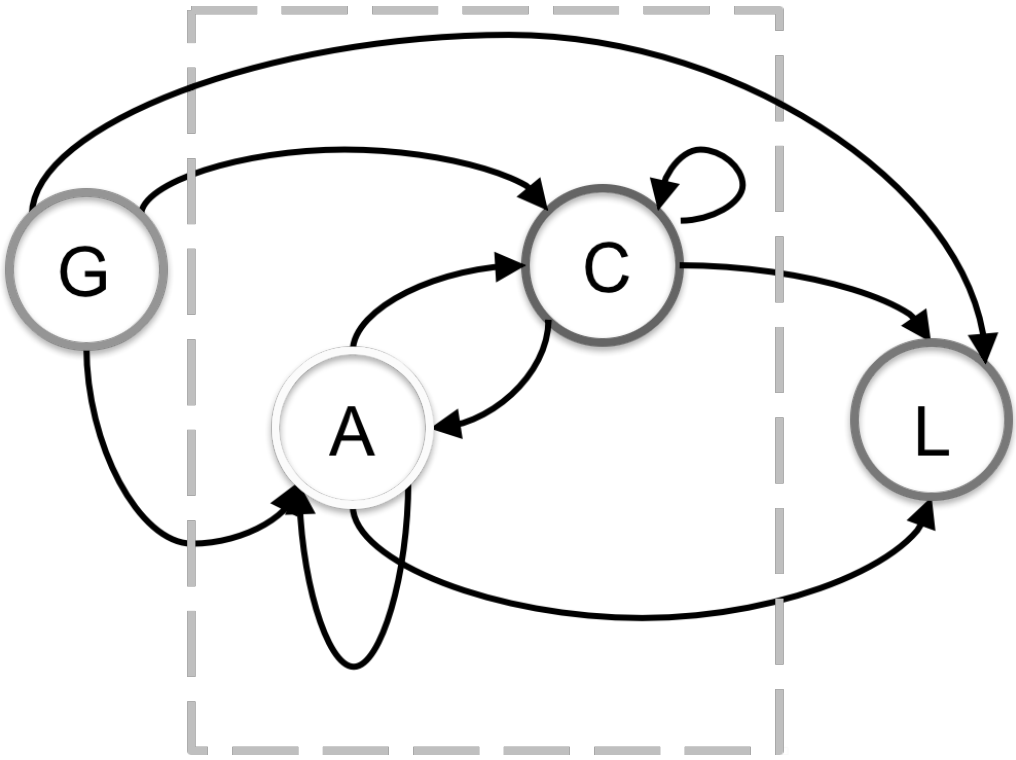


Figure 2.2 – A simplified view of the phases of casual talk described by Ventola - Greeting, Approach, Centre, and Leavetaking. Note that Approach and Centre phases may freely recur.

·	elements on either side are mobile around dot.
^	fixes order of surrounding elements
()	Enclosed element(s) are optional
$\widehat{\ }$	Elements below can be recursive.
[]	limits mobility. In following example ApD and ApI may switch places but both must precede the next element in sequence. $[\widehat{A}\cdot\widehat{ApI}\text{^}]$

Using this terminology she generated stings licensing conversations at different levels of size and depth (minimal to maximal) and between interlocutors at various levels of social distance (minmal to maximal). She gives an example of minimal conversation at minimal social distance (below). This conversation could be a quick chat or exchange between friends.

[G $(\cdot Ad)$^]{[$(\widehat{ApD})(\cdot\widehat{ApI})$^]}$(Lt)$^$(Gb)$

Her example of non-minimal conversation at minimal social distance is a deeper interaction with one or more centring elements (below)I.

[G $(\cdot Ad)$^]{[$(\widehat{ApD})(\cdot\widehat{ApI})\cdot\widehat{C}$^]$\widehat{Lt}$}^$Gb$

In her example of minimal conversation at maximal social distance, the conversation does not enter a centring phase.

(G) ^ { [$(\cdot Ad)$^][$(\widehat{ApD})(\cdot Id)$^]}$(Lt)$^$(Gb)$

For a non-minimal conversation at maximal social distance, she shows how strangers may enter a conversation and only introduce themselves after it has developed to centring stages.

(G) ^ { [$(\cdot Ad)$^][$\widehat{ApD}$^$\widehat{C}(\cdot Id)$^]$(\widehat{Lt})$}^Gb

Slade & Eggins' Chat and Chunk Phases

Slade and Eggins contend that casual talk can be seen as sequences of 'chat' and 'chunk' elements (Eggins and Slade, 2004, p. 230). Chunks are segments where (i) 'one speaker takes the floor and is allowed to dominate the conversation for an extended period', and (ii) the chunk appears to move through predictable stages – that is, it is generic. 'Chat' segments, on the other hand, are highly interactive and appear to be managed locally, unfolding move by move or turn by turn, and are thus amenable to Conversation Analysis style study. In a study of three hours of conversational data collected during coffee breaks in three different workplaces involving all-male, all-female, and mixed groups, Slade found that around fifty percent of all talk could be classified as chat, while the rest comprised longer form chunks from the following genres: storytelling, observation/comment, opinion, gossip, joke-telling and ridicule. Excluding ridicule and chat, which are not amenable to genre analysis, Table 2.1 shows the relative frequency of the encountered genres. Slade found differences in the distributions of the various genres between the differently gendered groups.

Slade and Eggins note that the chunk vs chat distinction could also be described as a cline of decreasing amenability to generic analysis, as shown below:

Text type	Percentage
Storytelling	43.4
Observation/Comment	19.75
Opinion	16.8
Gossip	13.8
Joke-telling	6.3

Table 2.1 – Relative frequencies of genres in Slade's workplace conversations

Genre	Generic Structure
Narrative	(Abstract) ^ (Orientation) ^ Complication ^ Evaluation ^ Resolution ^ (Coda)
Anecdote	(Abstract) ^ (Orientation) ^ Remarkable Event ^ Reaction ^ (Coda)
Exemplum	(Abstract) ^ (Orientation) ^ Incident ^ Interpretation ^ (Coda)
Recount	(Abstract) ^ Orientation ^ Record of Events ^ (Coda)
Observation/Comment	(Orientation) ^ Observation ^ Comment ^ (Coda) ^ (Completion)
Opinion	Opinion ^ Reaction ^ (Evidence) ^ (Resolution)
Gossip	Third Person Focus ^ Substantiating Behaviour ^ (Probe) ^ Perjorative Evaluation ^ (Defence) ^ (Response to Defence) ^ (Concession) ^ (Wrap-up)

Table 2.2 – Generic structure for 'chunk' types amenable to such analysis, () indicate optional stages

storytelling –> opinion –> gossip –> chat

They cite 'degree of interactivity of the genre and degree to which interpersonal or experiential meanings are foregrounded' as the factors influencing the adequacy of a generic description for each of the genres on the cline. They also provide a description of the stages or phases of each genre, sub-classifying stories as Narrative, Anecdote, Exemplum, or Recount:

Figure 2.3 shows two stretches drawn from the data where one contains more interactive chat the second more monologic chunks.

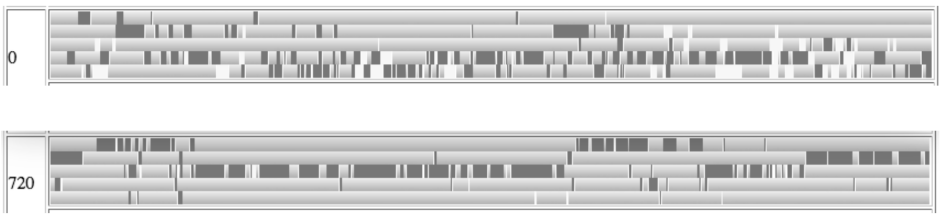


Figure 2.3 – Examples of chat (top) and chunk (bottom) phases in two stretches from a 5-party conversation in the D64 corpus. Each row denotes the activity of one speaker across 120 seconds. The numbers to the left denote the point (in seconds) during the conversation where the extract starts. Speech is dark grey, and laughter is white on a light 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.

The two descriptions of the structure of casual talk above differ in several respects. Ventola's phases span an entire conversation from greetings to goodbyes, whereas Slade and Eggin's chat and chunk phases do not explicitly treat the marginal phases. Ventola concentrates on the *function* of the phases, excluding centring, in creating and maintaining relations rather than on their *form*. Slade and Eggins stress the interactive nature of the chat phase, versus the more dominated nature of the chunks or long turns. In further analysis of the various genres encountered in conversational chunks, the function is considered, as this gives the characteristic form to each genre. Much of the work in this thesis examines the characteristics of chat and chunk phases of long casual conversations.

2.4 Conversational Floor Dynamics

From the above it can be seen that casual conversation, and indeed conversation in general, has been studied to a greater or lesser extent in a number of disciplines. Researchers have examined many facets of conversation including syntax and lexis, function, and prosodic and paralinguistic features. Much of the research has been based on transcripts of conversation, text analysis, or analysis of isolated prosodic

patterns, and corpus studies have largely been limited to collections of task-based interactions, while existing corpora of casual conversations have been restricted to relatively short interactions. In this thesis the focus is on the 'superstructure' of conversation - on the distribution of speech, silence, laughter and disfluency in long multiparty casual conversations. At this level, patterns of interaction can be seen without reference to lexicon, syntax or local prosodic effects. There has been interest in the timing of utterances in conversation and the overall patterning of vocalization and silence, or 'chronemics' of spoken interaction, from various disciplines over the years. Below, major research drives and important concepts are briefly described.

Conversation proceeds as a series of contributions by different participants. The pattern of these contributions varies according to the type of interaction, and a number of aspects of conversation can be inferred from analysis of the patterns of speech and silence observed in interaction. A major advantage of such analysis is that it is based on observance of the *presence* of speech, silence and sometimes laughter, and thus does not depend on human judgement of participant intent (Feldstein and Welkowitz, 1978)

2.4.1 Units of analysis for speech intervals

A logical unit for analysis of spoken conversation would seem to be a stretch of speech from a participant bounded by silence from that same participant. However, the definition of this unit has proven quite complicated. Much early work on the dynamics of speech and silence in conversation involved exploration of the various configurations of speech and silence possible in a two-way telephone conversation for the purposes of telephony engineering. In this domain, two definitions of a 'talkspurt' or stretch of speech from one party were proposed. The first, Norwine and Murphy's 1938 talkspurt, was defined as:

> 'speech by one party, including her pauses, preceded and followed, with

> or without intervening pauses, by speech of the other party perceptible to the one producing the talkspurt'.

(Norwine and Murphy, 1938)

This definition is problematic to operationalise, as it depends on perception by one of the interlocutors and also allows for the inclusion of within- or intra- speaker silences of undefined length. Brady, in his statistical analyses of 'on-off speech patterns' (Brady, 1968, 1969), reduced expanses of speech and silence to binary strings of 5 millisecond labels (0 for silence or 'off' and 1 for speech or 'on') where speech is automatically detected above a threshold. He defined a 'spurt' as an unbroken string of ones and a 'gap' as a string of zeroes, which together formed a 'spurt-gap' pattern such as 1001111... Brady noted that speech considered continuous by a human observer could contain short gaps 'due to stop consonants and slight hesitations', and thus suppressed silences of less than 200 milliseconds, and created two definitions which relied on a listener's perception.

The 'Brady' talkspurt was defined as:

> 'A talkspurt is a time period which is judged by a listener to contain a sequence of speech sounds unbroken by a pause.'

while a pause was defined as:

> 'a time period which is judged by a listener to be a period of non-talking, other than one caused by a stop consonant, a slight hesitation, or a short breath.'

(Brady, 1965, p4)

Although Brady manipulated the raw on-off data to more closely model human perception, the approach remains problematic, as 200ms is quite a long time in speech, and many pauses of this length may not be due to stop consonants.

Stretches of speech and silence by a single speaker have been referred to using a variety of terms including interpausal units (IPUs), utterances, and turns. A comprehensive account of the various terms is given in Edlund's thesis (Edlund, 2011), and this topic is returned to in the discussion on annotation in Chapter 4.

2.4.2 Speech and silence activity in conversation

Once speech and silence or on-off patterning have been annotated it is then possible to use these features to analyse spoken interaction. In an n-party conversation where each participant may be speaking or silent at any moment, there are n^2 possible states for the dialogue at any time, accounting for single party speech, overlap of one or more speakers, and global silence. Most early work on the dynamics of speech and silence in conversation involved exploration of the various configurations of speech and silence possible in a two-way or dyadic conversation. The 'on-off' data has been used to build stochastic models of dialog progression, and to compare different dialogs and interaction patterns along a range of dimensions. Conversations and psychiatric interviews were modelled by Jaffe and Feldstein, using data derived from a process similar to Brady's except for the sample length (300ms vs. 5ms). (Jaffe et al., 1964, 1967).

Jaffe and Feldstein's 1970 model of on and off talk was used in later work on dynamics of 'interpersonal timing' in phenomena including adult spoken interaction in stressed and unstressed conditions, and adult interaction with pre-linguistic infants (Cassotta et al., 1967; Beebe et al., 1988; Jaffe et al., 2001). The model, as described in their 1988 work on mother-infant play (Beebe et al., 1988) classified sound-silence patterns into five parameters - speaking turns, vocalizations, intrapersonal pauses, switching pauses and simultaneous speech. Turns were defined as the time between speaker switches, vocalizations as 'stretches of uninterrupted speech by the speaker who has the turn', interpersonal pauses rather confusingly referred to joint silences

bounded by the *same* speaker, while switching pauses referred to joint silences bound by vocalization of different speakers.

The distribution of silence and speech is fundamental to Sacks, Schegloff, and Jefferson's work on turntaking, where they described silences as pauses, gaps or lapses (Sacks et al., 1974). Pauses are bound by speech by the same speaker, gaps occur between speakers, while lapses refer to longer gaps where conversation is suspended for a time. Overlap refers to stretches of more than one speaker speaking at a time. In conversation, where there is no clear formal or role-dependent attribution of turns, turn taking has been described as locally organised. This local organisation and the equal speaker rights in force during casual talk pose interesting questions on how speakers know when to speak. There is considerable work on turntaking in dyadic and group situations, often based on Sacks, Schegloff, and Jefferson's seminal work on a turntaking algorithm which lays out the steps followed by participants in allotting turns. However, Sacks, Schegloff, and Jefferson pointed out in their work that this algorithm was not intended to cover all conversational situations. Identification of turn relevant places, junctures in conversation where it is possible for turn change to occur, has been based on reactive models - where speakers either react to perceived speech and respond (Duncan, 1972), or predictive models, where participants use cues including prosody and syntactic knowledge to predict or project the likely position of an upcoming TRP, and time their next contributions accordingly (Sacks et al., 1974).

Speech and silence distribution has also been important in work on interspeaker effects in interviews where co-ordination of and gap length between interlocutors has been demonstrated (Matarazzo et al., 1963, 1964).

Edlund and Heldner analysed pauses, gaps, and overlaps in three corpora of spontaneous speech in order to examine earlier claims of precision timing in conversation - particularly the claim that human conversation was so precisely timed that it tended

towards 'no-gap, no-overlap' (Heldner and Edlund, 2010). Their 2010 paper contains an exhaustive account of work on pause, gap and overlap distributions. In the current work, the terminology employed by Edlund and Heldner to describe conditions of silence and overlap in conversation is used.

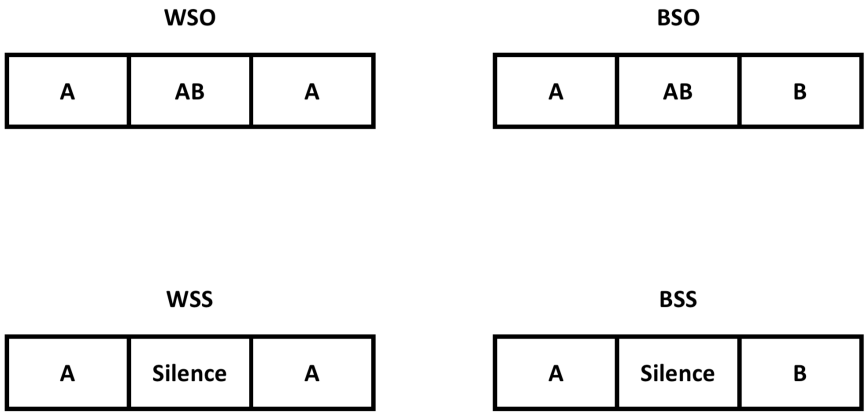


Figure 2.4 – Within and Between Speaker Silence (bottom) and Overlap (top) transitions

Their scheme defines four states of silence or overlap based on dyadic interaction as shown in Figure 2.4, the transition types for a speaker, A, are :

WSS Within Speaker Silence - A speaks before and after a silence.

BSS Between Speaker Silence - A speaks before silence, A is not speaking after silence.

WSO Within Speaker Overlap - A speaks before, during and after overlap.

BSO Between Speaker Overlap -A speaks before and during overlap, A is not speaking after overlap.

There has been less work on timing, silence and overlap in multiparty dialogue. The usefulness of speech/silence based analysis has been shown by Laskowski's work

on timing, silence and overlap in multiparty dialogue. He has used Markov models based on speech-silence patterns in multiparty meetings to analyse turntaking and overlap dynamics (Laskowski and Jin, 2010) and perform text-independent dialog act classification (Laskowski and Shriberg, 2009). He has also shown that analysis at the level of speech and silence dynamics can distinguish conversation types (Laskowski and Schultz, 2007), and infer participant characteristics including role, gender, and seniority (Laskowski and Schultz, 2008; Laskowski, 2014). Work in social signal processing has used speech dynamics measures such as overlap rate and speech/silence composition, often with prosodic features, to extract personality traits of speakers in dyadic and multiparty interaction (Ivanov et al., 2011; Gilpin et al., 2018)

In this thesis, measures of speech, silence, overlap and laughter are used to explore the structure of casual conversation, characterise its constituent chat and chunk phases, and explore the conditions around speaker change.

2.4.3 Interactivity in Conversation

There has been much interest in recent years in emergent properties of conversation related to the tenor or level of engagement experienced by participants. Relevant theoretical approaches have included notions of conversational temperature, rapport, involvement and engagement resulting in more 'engaged' segments or 'hotspots'. Several features have been proposed for use in modelling such engagement, generated from measures of phenomena including speech dynamics, prosody, eye-gaze, facial expression, gesture and posture (Ishii and Nakano, 2008; Jokinen, 2011; Oertel and Salvi, 2013; Müller et al., 2018). In addition to annotation of the relevant phenomena (eye-gaze, gesture, etc), such modelling relies on data previously annotated for engagement or involvement, which poses problems as these high level phenomena have not yet been definitively defined for operationalization. A simpler, but related conversational property is the 'interactivity' or 'liveliness' of a conversation. Interac-

tivity has been a subject of much interest in dyadic telephonic communication, as a measure of quality of experience, with much emphasis on the effects of delay or lag on perceived interactivity. Interactivity has been described in terms of conversational temperature – based on how much change occurs in conversational states, in terms of speaker change and interruption (Reichl and Hammer, 2004). This approach, and conversational traces or on-off patterns of speech activity have been used in building dynamic stochastic models of conversation based solely on speech, silence (and sometimes laughter) activity, resulting in models similar to those mentioned above, which predict conversational state on data quantized into uniform time steps. In this thesis, experiments are based on conversation or phase level static measurements, although a dynamic model is a clear avenue for later work. Therefore, interactivity is investigated using two static methods, *compression* (Burger et al., 2008) and *floor change rate*, a simple measure of floor/speaker change. Compression is briefly described below.

Conversations and dialogues have been compared in terms of the level of overlap of speech or laughter, which may point toward the level of interactivity or liveliness, by means of the notion of 'compression' . This measure, defined by Burger and Laskowski, was used by them to allow comparison of different speech exchange systems, such as meetings and seminars (Burger et al., 2008). Compression is defined as the total contributions by all participants in the conversation divided by the amount of time during the conversation devoted to these contributions – that is the total individual contributions without considering whether they are in the clear or overlapped (so each interval of overlapped speech is counted as many times as there are overlapping speakers) divided by the conversation time used for these contributions (so overlapped speech is counted only once) of the conversation.

Compression is used in this thesis to measure the level of overlap of speech and of laughter in the corpus as a whole, in individual conversations, and for individual

participants. It is also used to compare the level of overlap in chat and chunk phases. A second measure, floor change rate, is proposed and used to measure conversational liveliness as described in Chapter 8.

The work in this thesis aims to contrast the distribution of speech, silence and laughter in long multiparty casual conversation. It can be seen that analysis at the level of speech and silence can give insight to a number of facets of interaction, and the work described in this thesis will lay the groundwork for later stochastic analysis of casual talk, and indeed of the contrasts between task-based and social talk.

Below, three phenomena which can be measured in conversation - prosody, disfluency and laughter, and thus may help in describing and differentiating different types of spoken interaction, are briefly described.

2.5 Prosody in Casual Conversation

Although spoken interaction is multimodal, speech is primarily an acoustic phenomenon. The sounds that make up language are very well described at the segmental (vowel, consonant, syllable) level, and the suprasegmental or phrase level prosodic stress and intonation patterns of utterances produced in isolation have been well documented. However, the prosody of natural spoken interaction is not as fully understood. Conversation does not sound like a series of sentences read from a book. Speakers pause, repeat themselves, speed up and slow down, as well as varying their pitch, intonation, and loudness (amplitude). Talk is produced by individual speakers in overlap with others as well as 'in the clear'. Much of this activity serves the mutual construction of dialogue. Abercrombie drew attention to the difference between language as studied by linguistics, which he likened to 'spoken prose', and conversational language. He described how when reading aloud a piece of written prose, intonation is easily inferred from the text as

> 'the intonation.. adds little information. However, if the reader tries to

> read aloud a written conversation, they cannot tell what the original intonations were, as 'the intonations contribute more independently to the meaning'.

(Abercrombie, 1956, 6) Gumperz notes that rather than taking an interactional perspective, much phonetic work is based on a Saussurian view where language as made up of words, phrases, and sentences. In this view prosody is 'somehow derivative' and 'can be treated as an expressive overlay that supplements and modulates the more basic propositional content'. He postulates that participants in conversation use 'empirically detectable signs' or inference cues which evoke interpretive schemata or frames to situate or 'contextualize' the interaction, thereby co-constructing the conversation (Gumperz, 1992). The cues used encompass prosody, code-switching, and non-verbal cues such as posture, gesture, and gaze (Auer, 1988). Gumperz argues that contextualization occurs at three levels - conversation mangagement (turntaking, signaling of relevance), sequencing (cueing implicatures, disambiguating speaker intent), and framing (cueing inference as to the nature of the interaction). Selting and Couper-Kuhlen ascribe the neglect of prosody to the literary bias in linguistics, particularly in structuralist paradigms, combined with the historical dependence on written examples of language before the advent of technologies allowing audio and video recordings (Couper-Kuhlen and Selting, 2006). They describe earlier approaches to prosody as 'grammatical' where intonation was seen to act at the sentence level with falling tones inferring statements and rising tones inferring questions, and other 'tunes' corresponding to attitudinal modifications to the message.They highlight two difficulties in the 'grammatical' approach to prosody - co-varying lexico-syntax (attitude is mostly conveyed by the words and the prosody working together) and the iconicity of prosody (noting Bolinger's comment that 'intonation has more in common with gesture than with grammar' - i.e. prosody is not arbitrarily ascribed a 'meaning'). They propose an interactional perspective on prosody where language situated in

interaction 'suggests interpretations through a complex interaction of verbal forms with contextual and situational factors'. In their treatment of phonological research practice, Kelly and Local first define 'language events' and describe a marginal case - newsreading - before defining conversation, a contrasting central case, as

> 'an activity which is interrupted and responded to, which exhibits a wide range of tempo and other uttered effects, which is accompanied by a wide range of gesture, and which is face to face and maximally interactive, which is spontaneous and has no source in written material.'

(Kelly and Local, 1989, 12).

Later work by Local and Walker sets out a series of considerations for the analysis of Phonetics of Talk-in-Interaction (PTI) (Local and Walker, 2005; Ogden, 2012), a sub-discipline of Conversation Analysis. Conditions around phrase endings and turn changes have been studied by annotating relevant cases using schemes such as the Tones and Break Indices prosodic labelling system (Beckman and Hirschberg, 1994) and the Intonational Variation in English (IViE) labelling system (Grabe, 2001).

In this thesis, much of the work on conversation dynamics focusses on what happens at the end of a speaker's contribution. A natural next step would be to examine prosodic conditions around prosodic conditions around turn change and retention. A study was conducted on a section of the data (Gilmartin et al., 2018), but the methodology used was superseded by that described in chapter 8, and thus this work is not included in the thesis.

2.6 Disfluency in Casual Conversation

The term 'disfluency' has been applied to a range of phenomena encountered in spoken language, including filled and unfilled pauses, repetitions, unfinished phrases, substitutions and insertions. While the term can be used to describe pathological

conditions of speech such as stuttering or stammering, here 'disfluency' is used to refer to disfluencies produced during normal speech. Disfluencies in speech - hesitations, filled pauses, repetitions and repairs - have often been accounted for as markers of planning in action, particularly as they tend to occur near the beginning of utterances (Goldman-Eisler, 1968; Beattie, 1983). Other theories of disfluency posit that some pauses and hesitations may be a strategy to highlight important new information, particularly when used when the speaker is informing the listener - in lectures, for example.

Fox-Tree defines disfluencies as

> 'phenomena that interrupt the flow of speech and do not add propositional content to an utterance. This includes long pauses, repeated words or phrases, restarted sentences, and the fillers *uh* and *um*.'

(Tree, 1995, 709).

In her studies of disfluency (Shriberg, 1994), Shriberg restricted her work to within sentence disfluency, providing an operational definition based on disfluency research in the field of automatic speech recognition, where the objective is to transform text containing disfluencies to a 'desired' fluent form by removal of disfluent material. She defined the disfluencies in her analysis of spoken corpora as 'cases in which a contiguous stretch of linguistic material must be deleted to arrive at the the sequence the speaker 'intended' (Shriberg, 1994, 1), giving the following example of an utterance containing three disfluencies where the contiguous material to be deleted is struck through with a horizontal line (for processing purposes, punctuation and capitalization were removed by Shriberg):

> ~~they~~ they basically reviewed ~~oregon's plan or~~ the oregon plan toward ~~uh~~ nationalizing health care

These disfluent stretches are defined in terms of three constituent parts, and an

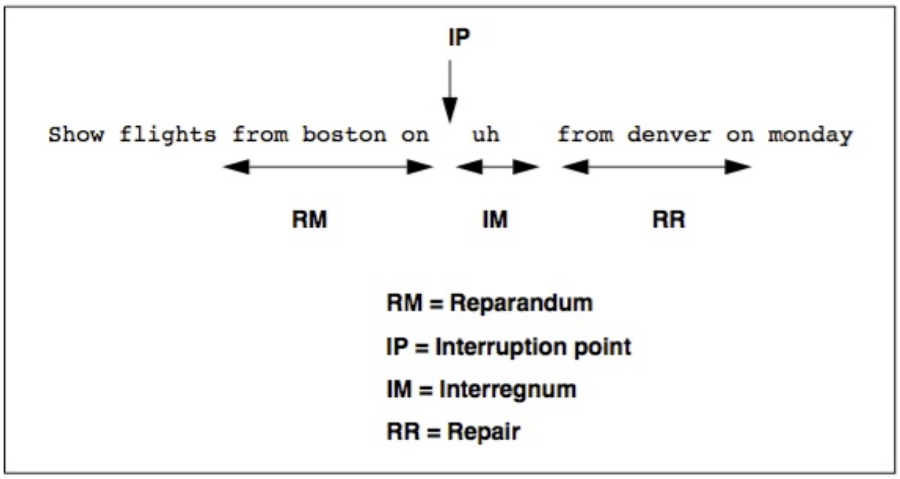


Figure 2.5 – Shriberg's illustration of the reparandum, interruption point, interregnum and repair constituents of a within sentence disfluency.

optional region. The **reparandum** (RM) consists of the words to be deleted. The **interruption point** (IP) refers to the point at the end of the reparandum where the interruption of a stretch of fluent spoken output is considered to take place. The **interregnum** (IR) refers to the region between the interruption point and the beginning of the **repair** section. The interregnum may contain editing terms such as 'I mean', filled or unfilled pauses or indeed be empty. The **repair** section (RR) consists of the stretch of new linguistic material which replaces the reparandum. In this view of disfluencies, the contents of the reparandum and the interregnum are deleted to arrive at the speakers 'intended' utterance. Fig 2.5, taken from Shriberg, illustrates these terms and the structure of a disfluency:

2.6.1 Frequency of Disfluency in Spontaneous Speech

In her 1995 study (Tree, 1995), Fox-Tree surveyed seven prior studies and found estimates of the frequency of non-pathological disfluency in speech ranging from 2 to 26 disfluencies per 100 words. She concluded that the variance was largely due to the inclusion or exclusion of unfilled pauses (short silences), and estimated that the

frequency of non-pause disfluencies in spontaneous speech is around 6 disfluencies per 100 words or 6%. Shriberg noted that disfluencies accounted for up to 10% of words and over a third of utterances in natural conversation. In terms of the type of disfluency most commonly encountered Shriberg reports that the highest rates of disfluency in her work were seen for filled pauses (FP), repetitions (REP) and deletions (DEL), which made up 85% of the disfluency tokens found in the Switchboard corpus.

2.6.2 Taxonomies and annotation schemes

One of the earliest taxonomies of disfluency, used to classify pathological disfluencies in stuttering adults and non-pathological difluencies of the average adult, was created and refined by Johnson over the course of a series of studies spanning several decades in the mid 20th century. The eight categories listed in his 1961 paper on reading and speaking rates in stutterers and non-stutterers (Johnson, 1961) have been widely used as the basis for disfluency research (Eklund, 2004).

The categories were:

1. Interjections of sounds, syllables, words or phrases. This category included 'extraneous' sounds such as uh, er and hmmm, corresponding to the filled pause of later research.

2. Part-word repetitions.

3. Word-repetitions.

4. Phrase repetitions.

5. Revisions, i.e. instances in which the content of a phrase is modified, or in which there is grammatical modification. This category also included change of pronunciation.

6. Incomplete phrases.

7. Broken words.

8. Prolonged sounds.

(Johnson, 1961)

There have been several taxonomies proposed for non-clinical disfluencies, most based on some form of Johnson's list. A full review of these is outside the scope of this thesis – comprehensive reviews of taxonomies can be found in Shriberg, Eklund, and Finlayson's respective theses (Shriberg, 1994; Eklund, 2004; Finlayson, 2014). Below, a fuller description of Shriberg's system is given, which is used in the annotation and analysis of disfluency in Chapter 7.

Shriberg's Pattern Labelling System

Shriberg's Pattern Labelling System for annotation of disfluency was developed for her thesis work (Shriberg, 1994). Based on an earlier system used at the Stanford Research Institute (SRI), the PLS has been in widespread use since. Similar schemes were used by Lickley in annotation of the MapTask Corpus for disfluency (Lickley, 1998), and by Eklund in his thesis work on disfluency in Swedish (Eklund, 2004). Shriberg's scheme was designed to be pre-theoretical and to allow disfluency information to be recovered from text transcriptions, with the ultimate aim of improving speech recognition results. The scheme is based on the linguistic sentence, and treats only within-utterance disfluencies. The core of the system is a 'Pattern Labelling System' which tags each element of a disfluency with a letter or symbol. Shriberg notes that in general, classification systems for disfluency are based on correspondences in wording between material in the RM and in the RR, and although different systems contain different numbers of classes, from as few as two basic classes to over a dozen, most distinguish at least the four classes of disfluency below, although exact terminology varies between systems:

1. correspondence but no change from RM to RR – as in repetitions

2. correspondence with modification from RM to RR – replacements and insertions

3. abandonment of material with no RR – fresh starts

4. no material in RM – filled pauses

Shriberg's taxonomy, using terminology adopted from Levelt's earlier work (Levelt, 1983), comprises six disfluency types - filled pause (FP), repetition (REP), Deletion (DEL), substitution (SB), insertion (IS) and articulation error (AC), as shown in Table 2.3. These tags then form patterns representing the disfluency and allowing for easy extraction and counting of disfluencies with basic text-processing methods. In the scheme the beginning and end of disfluent sequences are marked by [and], while a dot (.) is used to mark the interruption point. Repeated, substituted, deleted, and inserted words are represented by **r**, **s**, **d**, and **i** respectively.

Disfluency Type	Symbol	Example
Filled Pause (FP)	**f**	**uh** - we live in Dallas
Repetition (REP)	**r**	all **the** - the tools
Deletion (DEL)	**d**	**it's** - I could get it where I work
Substitution	**s**	any **health** cover - any **health** insurance
insertion	**i**	and I felt - I **also** felt
articulation error		and [**pin**] - pistachio nuts

Table 2.3 – Shriberg's Disfluency Labelling Scheme

The following examples from Shriberg illustrate the use of scheme:

are you you going to the party

[r . r]

when you we saw

[s . s]

he really wanted he wanted to see them

[r d r . r r]

I saw a car a yellow car

[rr . r i r]

did he want did she want

[r s r .r s r]

The scheme also includes diacritics which can be added to the element letters. Cut-off words are marked with a –, mispronounced words by a ~, and contractions by a ^.

The scheme includes tags for extra-syntactic and extra linguistic words. Extra syntactic words were words or phrases which did not contribute to the meaning of the sentence. This class includes explicit editing terms such as 'sorry' or 'I mean' appearing in the interregnum or editing phase which are tagged with **e** and other discourse markers such as 'well' or 'like' which are tagged **p**. The class also included filled pauses, tagged **f**, which can appear anywhere in the stream. The three classes are treated differently - filled pauses are treated as disfluencies wherever they occur, explicit editing terms are treated as disfluent and deletable only if they occur within a disfluency, while the situation is undefined with discourse markers even if they occur within a disfluency as their status as part of the intended original utterance is theory-dependent.

As mentioned above, the scheme is largely based on an earlier scheme used at the Speech Research Institute (SRI), but has several additions, and differs in the provision of an indexing function. In the SRI system, the words comprising repair and substitution sequences were indexed as r1, r2, etc., to reflect any changes in order of words. The SRI system also stipulated one to one substitutions. In Shriberg's scheme the indices were not used, a move which removed the constraint on substitution of varying length strings as in the example:

does the united flight serve a sn(ack)- breakfast

[ss- . s]

where the single word **breakfast** substitutes the two words **a snack** (note that in the example the word sn(ack) was incomplete (marked with –).

PLS also added methods for annotating serial and complex disfluencies. Two or more disfluencies are considered serial if the last word of the first disfluency preceded the first word of the second, and so on. In the example below the first disfluency is a deletion of 'which', while the second is the repetition of 'how much'.

which how much how much does that cost
[d.] [r r .r r]

In complex disfluencies, the component disfluencies include overlapping words, as in:

he she she went
[R[s . s] . r]

Shriberg analysed complex disfluencies in a compositional manner, as nested disfluencies. Outer levels were annotated with a capital letter depicting the output of correction of an inner level disfluency. In the example, the resolved substitution **[s.s]** is marked with **R** to show that it enters to next disfluency (repetition - **she she**) as the first element of a repetition. The scheme also accounts for ambiguous and degenerate cases. Ambiguous cases were disfluencies where there were several analyses possible. Two conservative conventions handle these cases - delete fewer over more words, and assume less over more correspondence. There were three degenerate classes. The first group were cases where degradation of the signal or factors such as background noise made speech unintelligible. A second group handled cases where labelling as a DF seemed inappropriate as in parenthetical asides or cases involving final ellipsis. The third class involves cases outside the scope of the PLS including uncorrected

speech errors and prosodic repairs, and DFs crossing noncontiguous regions (which are, however, of interest to the current work). PLS also includes two correction operators - an RM Deletor which deletes all words in the reparandum, and an **f** and **e** Deletor which deletes filled pauses and explicit editing terms anywhere in the DF. These two operators together with the pattern symbols outlined above should allow any sentence containing disfluencies within the scope of the scheme to be 're-engineered' to a fluent utterance.

2.7 Laughter in Casual Conversation

Laughter is regarded as a basic human social signals, believed to have evolved from primate facial displays - the relaxed open mouth display or play face. Laughter has been described as predominantly a social rather than a solo activity, is universally present in humans, part of the 'universal human vocabulary', innate, instinctual, and inherited from primate ancestors (Burling, 2007; Provine, 2004). Various accounts of the role of laughter exist, ranging from a response to humour to a social cohesion or bonding mechanism used since our primate days. It has been suggested that laughter is often a co-operative mechanism which can provide clues to dialogue structure (Holt, 2010). It has been described as a social cohesion or bonding mechanism present in primates (Glenn, 2003; Mehu and Dunbar, 2008). Laughter episodes take a range of forms – from loud bouts to short, often quiet chuckles. It is multimodal, comprising a stereotyped exhalation of air from the mouth in conjunction with rhythmic head and body movement (Bachorowski and Owren, 1995; Black, 1984). However, while this stereotypical laugh is generally produced upon asking an informant to laugh, it has been shown to be only one of several manifestations of laughter present in social interaction, and often not the most prevalent (Trouvain, 2003). In conversation, laughter generally punctuates rather than interrupts speech, although it can occur within speech as smiled speech or speech laughs. In recent years, there has

been increasing research into the occurrence and modelling of laughter in interaction (Laskowski and Burger, 2007; Holt, 2016; Ginzburg et al., 2014). Conversation analysis has highlighted connections between laughter and topic change; many conversations in the Holt corpus of mostly two person telephone dialogues include laughter at topic closings (Holt, 2010). Laughter has been linked to topic closure in situations where one participant produces jokes or laughs, thus inviting others to join in, with this invitation open to refusal in the case where interlocutors continue speaking on the topic at hand (Jefferson, 1979). Holt (2010) suggests that laughter may be prevalent at topic changes because turns consisting only of laughter are 'backwards' looking, not adding to the last topic, and thus constituting a signal that the current topic has been exhausted and that the conversation is at a topic change relevant point. In this thesis, laughter is considered an integral part of conversation, and it is included in the analysis of conversation structure in Chapter 6.

2.8 Dialog Technology and Casual Conversation

Over the centuries humans have developed ways of communicating beyond face to face spoken interaction. Reading and writing are major technological achievements, using language to allow the preservation and transmission of information across great geographical and temporal intervals. However, reading and writing are fundamentally linguistic activities, and thus limited in what they transmit - while a written account can describe the propositional information conveyed orally it lacks the multimodality of face-to-face communication. Of course, in many situations, only propositional information is needed, and most early surviving examples of written language are accounts of transactions and legal proceedings, inventories, and instructions for accounting, as in the ancient Sumerian clay tablets written in cuniform and depicting beer rations for workers dating from 3100 BC displayed in the British Museum. With time human literature has pushed the limits of (written) language to allow the

novelist, poet, and indeed journalist to evoke emotion in the reader at a distance, but books now lag TV as the leisure medium of choice for most - in the US Bureau of Leisure Statistics American Time Use Survey for 2014, TV watching occupied the greatest swath of leisure time (an average of 2.8 hours per day) while reading occupied much less leisure time; from a maximum of one hour per weekend day for over 75's to a mere 8 minutes per weekend day for 15 year olds, although the explosion in the use of the internet and text messaging has increased exposure to and use of text. In addition, while conversation is available to all, literacy has long been the preserve of the few, and the written culture of several millennia and many different human civilizations has simply not been available as a resource to many people.

Human communications technology has often exhibited a pattern of invention to meet practical needs - writing serving the needs of commerce, administration, and lawmakers - only for people to then use these means to extend the range of their interpersonal conversation or dialogue. People have scratched grafitti, written personal letters, and sent telegrams to express more of the mixture of propositional and social information present in face-to-face conversation. Written technology has been extended or adapted to allow transmission of these social signals. Ancient Irish manuscripts have margin notes with remarks on the monk scribe's attitudes or emotions. Soon after the invention of Morse code the number 73 and later 88 was appropriated to mean 'love and kisses', while emoticons or 'smileys' were appearing in email correspondence by 1982. Indeed, emoticon placement in emails has been shown to pattern after the placement of social signals such as laughter in spoken interaction (Provine et al., 2007), punctuating rather than interrupting propositional content.

The major drawbacks of communications technology for the expression of the full spectrum of conversation have been twofold. The first has been the lack of synchrony - the first media were essentially asynchronous - a message was written and transmit-

ted and there was a wait for the receiver to respond. The second has been the limitations of the media themselves - the written word, however well manipulated, is only ever an abstraction of the bundle of audio and visual information present in spoken interaction. With advances in communications technology the temporal lag has been reduced - from days and weeks to transmit letters, to minutes and seconds for telegraphy and telephony, and now audio and video communication can occur in almost real time. The problem of modality is also shrinking, audiovisual technology now makes the transmission of visual and voice information possible, more closely approximating the information available to interlocutors who are physically co-present. A third problem, that of access, is also disappearing, as humans no longer have to be literate in order to use mass communications media. The availability of near synchronous communication made the creation of artificial dialog or dialog systems possible - first as interfaces to perform practical tasks. More recently, the advent of high quality real time speech synthesis and recognition has led to growing interest in systems which not only perform a practical task or transaction but also engage in casual social talk with users.

Spoken dialogue technology seeks to create applications where a spoken dialogue interface allows users to interact with a system through natural speech. Systems are designed to take different surface forms - from telephony based Interactive Voice Response (IVR) to large back-projected full-body avatars equipped with speech and gesture capabilities to interact with multiple users.

Spoken dialogue technology has concentrated largely on task-based dialogue for much of its history. Early researchers realised that creating systems capable of general conversation was an intractable problem, but that it was possible to create systems capable of well-defined goal-oriented interactions (the Practical Dialogue Hypothesis) (Allen et al., 2001b). Tasks can range from information retrieval - systems providing responses to user queries on travel timetables (Raux et al., 2006), to provi-

sion of services such as education, entertainment, health promotion or companionship - tutoring systems (Graesser et al., 2005; Granström, 2004; Litman and Silliman, 2004), chatbots (Wallace, 2009; Christian, 2011), medication reminders, monitoring and mentoring fitness programmes, and elder care systems (Bickmore et al., 2010; Vardoulakis et al., 2012). These modules may function as discrete systems or as part of larger 'always on' systems. The first spoken dialogue systems were essentially typed text interfaces with speech synthesis (TTS) and automatic speech recognition (ASR) bolted on. As such, the dialogue involved was very practical, technological rather than social in terms of language, and created to perform a clearly defined task - a classic example is a banking system allowing users to query their accounts and perform transactions by working through a predetermined dialogue. In this example it is clear that successful task completion depends on clear transmission and reception of information - numbers, commands, etc. Research work based on task-based dialogue has formed the bedrock for theoretical and technical advances (Raux et al., 2006). A major concern was, and remains, the quality of ASR, which restricted the design of dialogue to terse exchanges which elicited only responses in the user which were as short and to the point as possible. With improving ASR, systems have become more conversational, allowing the user more leeway. Many systems now open with "How may I help you?" (HMIHY) type questions, allowing the user to respond freely rather than going through a series of restricted questions seeking to fill particular information 'slots' necessary for task completion. However, the focus is still very much on linguistic information recognition. However, even in these situations the notion of 'co-presence' is important - for example, customers tend to prefer systems where the voice sounds pleasant (Cheepen, 1988). Many research applications and platforms in spoken dialogue technology research have been based on dyadic travel or leisure arrangements. The travel domain was used in the DARPA Communicator research programme in the early 2000s, while tourism, travel and restaurant enquiries have

been the domains used so far in the annual Dialogue State Tracking Challenge/Dialogue System Technology Challenge (DSTC) (Williams et al., 2013).

In terms of casual or social talk, there have been systems built which attempt to engage users in friendly chat as part of a larger task-based dialogue in order to research whether such talk enhances task success and/or user experience and retention (Bickmore et al., 2010; Kruijff-Korbayová et al., 2014), and also systems which are exclusively social in nature (Yu et al., 2015; Mattar and Wachsmuth, 2012). However, these systems tend toward short dyadic 'chat' exchanges and do not attempt longer conversations or multiparty talk. Similarly, the recent Alexa Prize research challenges, which have led to increased interest in social spoken dialogue systems, also concentrate on short sessions of dyadic social chat or smalltalk. In order to work towards systems which can perform as a speaker or listener in longer multiparty casual conversations, quantitative knowledge of how humans behave in such interactions is needed. In recent years, social or affective computing has grown into an important field of enquiry, and there is interest in building systems which address social as well as informational needs of users. With more social systems, the ability to provide 'human-like' conversation becomes vital, indeed the very idea of a companionship system entails a 'companion' who converses with a user rather than simply performing practical tasks. It is hoped that the work in this thesis will provide pointers to the modelling of more realistic (multiparty) casual conversation.

2.9 Conclusion

Dialogue takes place for many purposes, where the task-based or casual components may be more or less important. Indeed, there is a continuum of possible spoken interaction from completely phatic chat to content critical formal spoken communication - for example pleasantries between two people at a bus stop would fall near the chat end of the spectrum while air traffic control communication between pilot and

control tower would fall at the content critical end. A complete description of dialogue should include both transactional and interactional components. Interactional and transactional dialogue differ along several dimensions. Below, the differences between casual and task-based talk, and the features of casual talk are briefly summarised.

Assuming all dialogue to be goal-motivated, the goal varies with the nature of the interaction. At the social end of the spectrum, the goal is to establish co-presence, build or maintain interpersonal bonds, entertain, build and maintain social identity, keep channels of communication open, and this goal may extend far beyond the life of an individual session or dialogue. In this type of dialogue participants may not be consciously aware of why they are chatting. In task-based dialogues, goals by definition are the accomplishment of the task, and are more discrete, short term, clearly defined and known to the participants. Contrast (i) a customer booking a flight at a travel agent's with (ii) a worker having a short chat with a workmate at the watercooler. At the travel agent's, the goal is a successful business transaction. Both participants are aware of this goal, and the task will typically succeed or fail in the course of one interaction. At the watercooler, the goal is to maintain a positive social relationship, and participants probably do not have this goal explicitly in mind, and the relationship continues over an extended period of time. The length of dialogue also varies, with task termination governing how long a transaction lasts, while casual talk can continue indefinitely. Interaction content varies with dialogue type. Small talk or chat between strangers or acquaintances is light, without controversial content, and generally positive, although with greater familiarity between interlocutors topics and weight of discussion can be 'heavier', as, for example, in 'dinner party conversation'.

Where information transfer or task completion is the object of conversation, content is related to the task, and there is often a mechanism to move the task along – in business meetings there is often an agenda to be followed and a chair to move the

meeting from item to item. In many transactional exchanges, what is spoken about is dictated by the transaction taking place. Many analysts use the notion of frames as a useful construct in understanding what happens in particular interactions. Therefore a SHOPPING frame would include expectations of questions about price, properties of the goods, a 'pitch' by the salesperson, etc. In social talk, there are no predetermined topics for the interaction. However, depending on the relationship among participants, there are topics which are more suitable, and topics which are avoided.

There are also differences in participant roles in transactional versus interactional talk. While transactional encounters (tutorials, medical consultations, job interviews) have institutional roles such as tutor/student, doctor/patient, interviewer/interviewee defined with fixed status agreed by convention, interactional encounters such as casual chat do not involve such fixed roles, and speaker rights may be viewed as equally distributed.

As previously mentioned, real life dialogues do not fall completely into chat or task-based categories, but move up and down the spectrum as necessary, throughout the duration of the interaction. For instance, in a business meeting, important exchanges where content is critical are interspersed with lighter intervals where good relations are maintained or reaffirmed. Conversely, office workers may be carrying on an intermittent chat throughout the course of the working day which provides the matrix in which more task-based dialogues are embedded. It is clear that understanding and indeed modelling of human spoken interaction entails knowledge of the mechanics of the various types of talk likely to be encountered.

Transactional talk has been far more widely studied than interactional talk, due both to a literary or instrumental bias in some communication studies and also to the lack of available data. However there are a number of studies which together give a picture of many of the facets of casual or social talk. Much existing work on social and casual conversation has concentrated on linguistic and discourse analysis of the con-

tent of these conversations, very often from tapescripts of dyadic conversations or indeed invented examples. Many researchers in the field identify the fact that they base their studies on orthographic transcriptions only as a limiting factor in the breath of their descriptions (Laver, 1975; Schneider, 1988; Thornbury and Slade, 2006).

Useful insights on spoken interaction have been gained by exploration of features which do not rely on lexical content. Such features, including speech patterning, distribution of speech and laughter and the parameters of silences and overlap have been used to infer differences in conversation genre, participant role, involvement or engagement, and to predict the course of turntaking. Another measurable characteristic of conversation is the occurrence of disfluency - a universal feature of spoken production, which may aid in differentiating different conversation phases.

The work in this thesis will add prosodic and temporal information, arising from studies of audiovisual recordings, to the existing body of knowledge describing multiparty casual and smalltalk. The explorations in the following chapters are corpus-based, describing the distribution of speech, silence, overlap, laughter, and disfluency in casual talk, and investigation whether these differ in the constituent chat and chunk phases of long form casual conversation.

The next chapter focusses on the data needed for this data-driven analysis of conversation, the challenges of obtaining casual conversation data, and the dataset used in the experiments described later in the thesis.

Part II

Methods: Data, Annotations, and Statistical Methods

Chapter 3

Data

"Perhaps the greatest single event in the history of linguistics was the invention of the tape recorder, which for the first time has captured natural conversation and made it accessible to systematic study"

- Halliday, 1994

3.1 Introduction

In this section the requirements for data suitable for the analysis of long form casual conversation are outlined, and issues in corpus collection are discussed. Available corpora are surveyed, and the three corpora, D64, DANS, and TableTalk, from which the casual conversations studied are drawn, are described. The process of preparation of the data from raw recordings is described. A brief description of the TEAMS corpus, used to contrast casual and task-based talk, is also given.

3.2 Conversation Data and Corpus Research

The information bundle in conversation includes audio (vocal) and visual signals and cues. Vocal information comprises verbal and non-verbal phenomena, includ-

ing speech (and silence), prosodic contours, voice quality, laughter, coughing, and aspiration noise. The visual channel contains gesture displays, posture variations, and other cues ranging from facial expression, hand and body movements, and eye-gaze to variations in skin colouring (blushing). Some phenomena are tightly linked to the linguistic content of utterances - examples include intonation, beat and iconic gestures, others contribute to the participants' estimation of attitudinal and affective information about each other.

In recent years the various facets of conversational exchange have been studied through corpus collection and analysis, with recordings capturing as much as possible of the signal bundle through multiple channels, including audio, video, and motion-capture. There has also been increasing interest in insights into affective and cognitive states gleaned from biosignals - data collected from the body including heart rate, electrodermal activity (also known as skin conductance, galvanic skin response), skin temperature, pupil dilation and blood pressure. With technological advances, such measurements can be made unobtrusively over prolonged periods of time using wearable sensors and monitors. There has also been interest in data collected from technology used during interaction such as mice, keyboards or shared whiteboards.

Although this study focusses on the audio channel of multiparty conversation, it was decided to use corpora with as many modalities recorded as possible, to allow future studies combining information from different modalities in analysis of conversation - as an example, the audio annotations described later in Chapter 4 could be combined with annotations of gesture or eye gaze drawn from video recordings to study multimodal aspects of turntaking. It should also be noted that multiparty talk, even with near field microphones, contains very much overlap and it can be very helpful during annotation and analysis to refer to video to disambiguate audio from different speakers.

3.3 Collection of speech and dialogue data

Early linguistic fieldwork relied on careful written transcription and, later, audio recordings of speech, where a phrase or structure was often elicited directly from an informant from the speech community of interest (Dalton and Ní Chasaide, 2003). Later research has used specially designed corpora created in various ways for different purposes, from read or acted speech recordings made in low-noise environments such as anechoic chambers for fine phonetic or prosodic analysis (Burkhardt et al., 2005), to tape recording of participants during their normal daily lives (Du Bois et al., 2000). These elicited pieces of language have informed surveys of phonological, syntactic, and semantic aspects of language and dialect.

To gather data on spoken interaction, speech has frequently been elicited by having subjects perform a task or take part in interactions with a clear short-term goal to elicit natural or spontaneous spoken interaction. It should be noted that in this field 'spontaneous' simply means unscripted. These 'task-based' interactions are often information gap activities completed by participants exchanging verbal information. Recordings of these tasks then form corpora on which study of dialogue is based. Two popular paradigms, those used in the HCRC Map Task corpus (Anderson et al., 1991) and in the LUCID (Baker and Hazan, 2010, 2011) and Wildcat corpora (Van Engen et al., 2010), will serve as examples.

The HCRC Map Task corpus is a set of audio recordings of 128 elicited unscripted dialogues, where English speaking participants performed a dyadic information gap exercise of a kind often used in second language teaching to elicit spoken interaction (Watcyn-Jones, 1981). The task involved two participants sitting in a room together without visual contact, with maps on which some features were incomplete in one drawing or another. One participant (the Giver) gave instructions on a route through the map which the other was to trace on his map; the differences in maps became talking points in the resulting dialogues. The corpus has been used widely in several

areas of speech and dialogue research, including turntaking mechanisms (Bull and Aylett, 1998), interspeaker effects such as priming (Reitter et al., 2006), and the distribution of silence and overlap (Heldner and Edlund, 2010),

The LUCID and Wildcat corpora both used the DiaPix task, where participants were given pictures and asked to find the differences by speaking about their pictures without seeing their partner's picture. The DiaPix task was created for the Wildcat corpus recordings. The main purpose of the task is to elicit talk from both participants equally, and thus roles were equal, in contrast to the Map Task where one participant gave instructions or information and the other followed. In the Wildcat corpus the task was performed by pairs with similar or differing first languages in order to study task completion rates where participants had differing linguistic backgrounds, while in the LUCID corpus the task was performed by dyads in the presence of communication hampering conditions such as background noise and hearing impairment. Other typical tasks used to gather spontaneous talk include discussions elicited by asking participants to make decisions together - a common paradigm is ranking items on a list - for example to decide which items would be most useful in an emergency (Vinciarelli et al., 2012). Games, where participants speak to each other as part of the gameplay or indeed while they are playing a game, provide an interesting source of often multiparty conversation data, examples include the TEAMS corpus (Litman et al., 2016) and the Sheffield War Games corpus (Liu et al., 2016)

There have also been large research projects devoted to collection of corpora of multiparty meetings. These have included real and staged business meetings as in the ICSI (Janin et al., 2003), and AMI corpora (McCowan et al., 2005). These corpora have been the source of data for much of the corpus research into dialogue in recent years.

The ICSI corpus, recorded at the International Computer Science Institute (ICSI) in the early 2000s, comprises approximately 72 hours of close and far field audio

recordings of 75 informal meetings held by various research groups at ICSI. A major feature of this corpus is that these meetings were not staged and would have taken place anyway. The meetings ranged from 3 to 10 participants, from a total of 53 speakers including native English speakers and non-native speakers of various levels of competence. Several speakers appeared in more than one meeting. The corpus and ensuing annotations have been used for research into a wide range of topics related to spoken dialogue and to speech technology, among them dialogue structure and dialogue acts (Ang et al., 2002), speaker overlap and ASR errors (Cetin and Shriberg, 2006), and stochastic analysis of turntaking styles (Laskowski, 2016), laughter (Laskowski and Burger, 2007) and genre (Laskowski and Schultz, 2007).

The AMI meetings corpus was published in 2006 as the result of several years of international collaboration in the AMI Consortium, and consists of 100 hours of predominantly role-played meetings held between mostly non-native English speakers. The corpus contains audio and video recordings as well as digital pen, whiteboard, and presentation slide data taken from the meetings. The corpus has been extensively annotated and has been used in studies of several aspects of spoken interaction, including detection of dominance (Aran et al., 2010), personality traits (Valente et al., 2012), and social role recognition (Sapru and Bourlard, 2014)

The speech in these task-based and meeting resources is spontaneous and interactive and such corpora have contributed greatly to our understanding of facets of spoken interaction such as timing, turntaking, and dialogue architecture. However, the interactions cannot be considered casual talk, and the results obtained from their analysis may not transfer to the less studied 'unmarked' case of casual conversation. In addition, the participants' motivation for taking part in an artificial task is not clear, and the activity is removed from a natural everyday context. It is not certain that tasks such as these can be used to make generalisations about natural conversation (Lemke, 2012), and they are certainly not examples of social casual conversation.

Corpora of natural spoken interaction have been collected by researchers in the ethnomethological and conversational analysis traditions by recording real life everyday home or work situations. Early studies used recordings of telephone calls, as in the suicide hotline data collected by Sacks, Schelgoff's emergency services data and the conversational data collected by Jefferson (c.f. for example (Sacks et al., 1974)), or recordings of televised political interviews (Beattie, 1983).

Extensive corpora of telephonic speech data have also been gathered by recording large numbers of phone conversations, as in the Switchboard corpus (Godfrey et al., 1992), the ESP-C collection of telephone conversations (Campbell, 2007), and the Callhome corpus (Canavan et al., 1997). In these corpora, subjects were paid to talk to one another, either following topic suggestions given by researchers or about whatever they wanted to discuss. While telephone corpora have the advantage of allowing very good capture of the audio signal and can capture more causal interaction, telephone calls may well constitute separate speech-exchange systems with characteristics differing from those of face-to-face talk. As this study focusses on multiparty speech, telephone corpora were not considered.

Audio corpora of non-telephonic spoken interaction include the Santa Barbara Corpus (Du Bois et al., 2000), sections of the ICE corpora (Greenbaum, 1991) and of the British National Corpus (BNC Consortium, 2000).

The strategy of surreptiously or openly recording speakers talking 'in the wild', or indeed giving participants recording equipment over extended periods of time to record natural conversation has been employed since the early days of Conversational Analysis in the 1970s and produces very natural data; but problems include the quality of the recordings, and the fact that phenomena of interest to researchers may not occur frequently enough in the data to merit the expense of creating the corpora, not to mention ethical concerns in the case of surreptitious recordings (Warren, 2006). Examples of recordings of casual talk in the wild include Warren's surreptitious

recordings of conversations overheard in public places such as restaurants (Warren, 2006), Slade's recordings of workplace conversations (Slade, 2007), Wilson's recordings of Belfast adolescents (Wilson, 1989), and recordings where researchers started conversations with strangers in public places such as the audio collected by Schneider and Ventola for their respective theses (Schneider, 1988; Ventola, 1979).

More recent high quality recordings of everyday talk have been obtained by having subjects wear a microphone throughout the course of their daily lives for extended periods (Campbell, 2006), while others have amassed audio and/or video collections of recordings of different types of human activity, as in the Gothenburg Corpus (Allwood et al., 2000). There have also been large scale collections of substantial casual conversations (c. 20 minutes per recording) in the Nijmegen corpora of casual Dutch (Ernestus, 2000), French (Torreira et al., 2010), Spanish (Torreira and Ernestus, 2012) and Czech (Ernestus et al., 2014).

With increasing access to video recording and motion capture equipment, and awareness of the multimodal nature of human spoken interaction, rich multimodal corpora are appearing, comprising naturalistic encounters with no prescribed task or subject of discussion imposed on participants. 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), as well as extensions to the Gothenburg Corpus, and the English language Cardiff Conversational Database (Aubrey et al., 2013). Many of these corpora are based on dyadic conversation and the interactions are quite short, ranging from 5 to 20 minutes in duration. Thus, while they are very valuable for the study of dyadic interaction, particularly at the opening and early stages of interaction, they are not ideally suitable for analysis of multiparty longer conversations.

Earlier corpora such as the British National Corpus and the Gothenburg Corpus often consisted of collections of examples of exchanges collected opportunistically or

by arrangement. Such collections pose a challenge for corpus studies in that comparison between different examples is not simple, and can be made more difficult due to differences in recording methods and conditions. Newer corpora are often designed to address these problems by collecting several examples of interactions of similar duration, carried out in the same or very similar conditions, with balance wherever possible in participant numbers and gender, allowing for ease of statistical analysis.

In the next section, issues impacting the suitability of corpora for the work in this thesis are briefly discussed and the requirements for data are outlined.

3.4 Issues in Corpus Collection

Collecting appropriate data for the study of human spoken interaction is a difficult task, with several trade-offs to be made between the need to obtain clear recordings which can be analysed reliably using existing analysis tools (experimental control), and the need to record real, or at least realistic, conversational data (ecological validity). These often opposing imperatives lead to important choices during corpus creation and in the selection of a suitable corpus for analysis. Below is a brief outline of the most important of these issues - environment or setting for recordings, recording equipment and modalities used, and nature of task or interaction type, and an additional consideration in corpus choice - the existence of annotations / other analyses of the data in the corpus.

3.4.1 Setting and Participants

Ecological validity is often a problem when attempting to create models of real-life interaction from corpora. To capture genuine human interaction, the data should be collected in an environment natural to the type of interaction. Ideally, recordings would take place 'in the wild', but such recordings may not be of very high quality and recording conditions are not consistent. To ensure sufficient audio and video

data of acceptable quality for a wide variety of analyses (acoustic studies of pitch, affect studies based on video analysis of facial expression, accurate automatic speech recognition), many recordings are made in artificial environments. In this situation, creation of a suitable setting is paramount. For example, a boardroom setup may be more appropriate for collecting meeting-type interaction, while sofas and armchairs or a dining setting would be better for casual conversation. However, the presence of a number of speakers in the same room means that audio separation of different voices can be poor, and so to ensure good channel separation recordings are sometimes made with participants communicating remotely from separate rooms or even soundproof booths. This situation is obviously artificial and may not result in natural interaction. The issue of ecological validity can also affect the ease with which experiments can be controlled. In the case of naturally occurring talk, it is very difficult to get data 'to order' and thus participant-related factors such as differences in the age, social and educational background, personality and gender could confound results of analyses, particularly when the dataset is not large enough to even out such differences. However, controlling for these factors by recruiting similar participants may decrease the chance of recording the most natural data possible. Participant differences and indeed environmental differences such as time of day, season, and even world events could also make comparing different corpora difficult and affect the reliability of results.

3.4.2 Recording equipment

Most human interaction is face to face, and as visual information forms a large part of the signal bundle, video should be taken of the interaction from as many angles as possible. To obtain clear audio signals, multiple microphones are necessary, with the ideal being one close-coupled microphone per participant. With increasing interest in ubiquitous computing and an ever expanding range of affordable and non-invasive

sensing devices it is becoming more and more common and useful for recordings to be made of biosignals such as heart rate and galvanic skin response. However, the more equipment present in the setting, the less close to a natural environment the setup for participants.

3.4.3 Nature of interaction

While much progress has been made, it is not certain that one size fits all for interactional data, and that results and insights gained, for example, from measurements of the timing of pauses and gaps in telephone conversations about an artificial item ranking task or indeed in real-life suicide hotline recordings, are generalizable to face-to-face casual conversation. On the other hand, 'ad hoc' collections of real interactions may not have suitable recording conditions or indeed recording formats for analysis, although some older collections can be upgraded using digital technology, as with the British National Corpus (Coleman et al., 2011).

3.5 Choice of Corpora

A survey of corpora of spoken interaction was made, resulting in a list of corpora which were widely available, contained 'spontaneous' spoken interaction, and were in English. The choice of language was dictated by the human resources available and nature of the annotation task - the author would be the sole annotator for many of the data, and the chat and chunk annotation is based on a human description of the kind of talk encountered in the two conditions. These corpora were compared in terms of content (natural conversation or task dependent), setting, recordings, nature of interaction, existing analyses, and ease of access in terms of financial cost (see Table3.1). Several corpora were excluded at this stage including the Switchboard corpus (dyadic, audio only, telephonic speech only), Santa Barbara Corpus (audio only), and HCRC Map Task (dyadic, audio only).

Corpus	Setting	Range of media	Existing analyses	Spontaneous	Ease of Access	Type
AMI	Meeting Room	AV	many	Y	Y	multiparty
Cardiff Conversational Database CCDb	Lab	AV	some	Y	Y	dyadic
D64	Lab Living Room	AV	some	Y	Y	multiparty
Dans	Lab Living Room	AV	some	Y	Y	multiparty
ICSI	Meeting Room	A	many	Y	Y	multiparty
MapTask	Lab	A	many	Y	Y	dyadic
TableTalk	Lab Table	AV	some	Y	Y	multiparty
Switchboard	Phone	A	many	Y	Y	dyadic

Table 3.1 – Comparison of Conversational Corpora, for range of media A = audio and V = video

It was decided to focus on corpora recorded with as many media as possible. This will allow for further experimentation to link cues other than audio to the insights gained in the current work. The D64 (Oertel et al., 2013) has the advantage of suitability, as the corpus consists of casual conversations containing both light chat and more propositional information rich dialogue (chunks), in a natural setting. It is very richly recorded with audio for individual speakers and 5 camera angles. The only disadvantage is that all annotations would need to be done by the author. The DANS corpus (Hennig et al., 2014) also contains suitable casual talk, but does not have as many recording sources (microphones) as the D64. However, it has been partially annotated for topic shift, laughter, speaker, and perception of engagement, but would require review and transcription. The TableTalk corpus (Campbell, 2009) also contains spontaneous talk and is video recorded but the transcriptions were found to lack accuracy and thus would need to be redone. The AMI corpus and ICSI corpora have the advantages of being very richly recorded and previously transcribed and annotated, and are large enough to allow robust statistical analysis. However, the interactions were meetings rather than casual talk although they did contain short stretches of casual talk. In addition, in the AMI corpus, dialogues are often roleplays and therefore do not score highly for ecological validity. It was decided to use the D64, TableTalk and DANS corpora. As there were differences in annotation between them and large amounts of unprocessed data, it was decided to segment/re-segment all of the recordings used and to annotate all of the conversations of interest using a common scheme as described in Chapter 4. A brief description of each of the three corpora is given below.

3.6 Corpora Used in Analysis

In this section, the three corpora used in the following chapters are described. Details are given of setting, participants, recordings, and preliminary preparations made for each corpus to synchronise these large data collections for annotation and analysis.

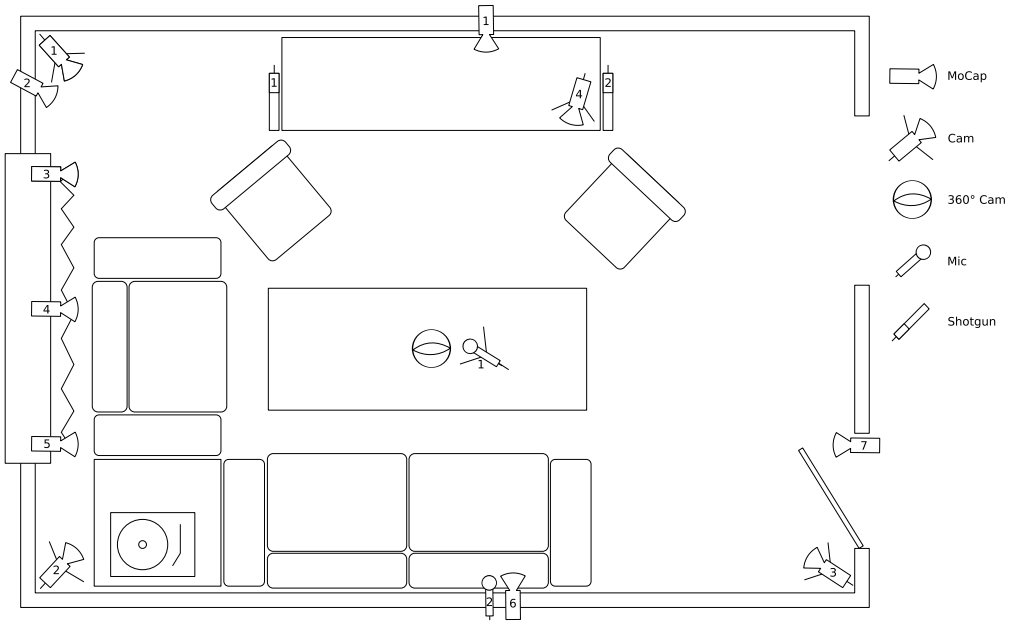


Figure 3.1 – Schematic of the setup for recording in the D64 corpus

3.6.1 D64 Corpus

The D64 corpus is a multimodal corpus of over 8 hours of informal conversational English recorded in three sessions in Dublin in 2009. The setting was an apartment living room, arranged as shown in Fig. 3.1, which also shows the distribution of recording equipment for the first recording session. Several streams of video, audio, and motion capture data were recorded for the corpus. The objective was to capture the most natural conversation possible in the circumstances and thus 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. The design and collection of the corpus is fully described in (Oertel et al., 2013).

Sessions 1 and 2 were recorded in the morning and afternoon of the same day with session 3 recorded on the following day. Recordings were made on a range of cameras, microphones, and motion capture equipment, resulting in a large bank of

video and audio files of different lengths with varying start and stop times, in addition to several motion capture recordings.

The intervals of interest to the current research were those where the group sat around a coffee table talking together. These sections occurred in all three sessions, with 5 participants in the first session, 4 in the second, and 3 in the third.

D64 Corpus Organisation

The recordings generated over 50 Gb of video files, audio files, and motion capture files. The first step in data preparation was to create copies of the master recordings, organise these chronologically, devise and apply a naming protocol, and resample all audio recordings to 16Kb/s to make them more manageable for use with annotation software.

Recordings were named by session, recording device, and 'tape number' - as each camera filled its internal storage - so S1C3002 is a video from Session 1 recorded on camera 3, and is the second of the camera 3 tapes for that session. Audio files were named for the microphone used (S1 = shotgun1) or for the participant wearing them in the case of chest microphones. All information was stored on a Microsoft Access database.

As the original recording file sizes were too large for smooth use of annotation software, the video and audio files were down-sampled to 720p (1280x720 pixels) and 16Kb (16,000 samples per second) respectively, levels which would provide sufficient quality for annotation purposes and be much less resource-heavy for processing. This also allowed the creation of a more easily distributable 'compact' corpus.

Participants

All participants were native speakers of English or spoke English in their work lives and had at least a C1 level of English on the Common European Scale (Little, 2006),

and thus were near-native or functionally bilingual.

Name	Sex	Age	First Language	Country of Origin	Other Languages
P1	M	60's	English	UK	Japanese
P2	M	50's	Swedish	Sweden	English
P3	M	50's	English	Ireland	German
P4	F	20's	German	Germany	English
P5	F	20's	Dutch	The Netherlands	English

Table 3.2 – D64 Participants

Table 3.2 gives details of participant demographics. The participants were not strangers. P1 knew P2, P3, and P4 before the recordings, P3 and P4 were acquaintances, P2 and P4 knew one another, and P4 and P5 were friends.

3.6.2 DANS Corpus

The Dublin Autonomic Nervous System (DANS) corpus is a multimodal corpus of approximately hour-long informal English conversation between several participants recorded in a living-room setup at the Speech Communications Lab, Trinity College Dublin in 2012.

Figure 3.2 shows the setup and participants chatting on Day 2. Again participants were encouraged to talk or not as the mood took them.

In addition to the audio and visual recordings, participants wore Q-sensors, wristbands which measure the electrodermal activity (EDA) or galvanic skin response in the wrist - the level of conductance based on the level of perspiration on the skin. EDA has been linked to levels of emotional arousal (Dawson et al., 2007) and to cognitive load (Shi et al., 2007), and thus the EDA data may be very useful in future in extraction of important or content-rich sections of dialogue but were not used in the work described in this thesis.

Corpus Description

The DANS corpus comprises three sessions of informal English conversation recorded over three days in a living-room like setting with sofa and armchair.

Name	Sex	Age	First Language	Country of Origin	Other Languages
P1	M	60's	English	UK	Japanese
P2	M	50's	English	Ireland	German
P3	M	30's	English	Ireland	Irish
P4	F	30's	English	US	Italian
P5	F	30's	French	France	English

Table 3.3 – Dans Participants

There were five participants in total - three male and two female, as shown in Table 3.3. Two of the males were Irish native English speakers, while the third was a British native English speaker. One female was a North American native English speaker while the other female was a French national who had lived in Ireland for some years. She worked in English and her spoken and written command of the language was at the C2 or near-native level on the Common European Scale (Little, 2006). Participants were free to speak about any topic they liked and to move around as they wished. P1 knew all participants previously, P2 knew P3 and P4, and P3, P4, and P5 knew one another. Session 1 consisted of informal chat between the American woman (P4) and the British man (P1). Session 2 contains three hours of conversation between the American woman (P4), the British man (P1), and one of the Irish men (P2). Session 3 contains informal chat between the American woman (P4), one of the Irish men (P3), and the French woman (P5). The camera setup for this session is shown in Figure 3.2. There was a camera focused on the sofa giving a clear view of speaker P4 sitting on left of sofa and speaker P3 sitting on the right of the sofa. Another camera gave a clear view of speaker P5 sitting on the armchair. Sound was recorded using microphones close to each speaker with an additional omnidirectional microphone on the coffee table between them. The corpus includes measurements of par-

Figure 3.2 – Three participants chatting on Day 2 of Dans data collection.

ticipants' electro-dermal activity (EDA), also known as skin conductance or galvanic skin response. This is a measure of perspiration in the skin, which is linked to arousal. All participants wore biological sensors (Q-sensors) on the underside of each wrist to measure EDA. Q-sensors (Poh et al., 2010) are wristbands which unobtrusively measure electro-dermal activity (galvanic skin response) in the wrist as well as three degrees of acceleration. While the wrist, in comparison to the palm or fingers, is a less sensitive location for recording EDA, these sensors were selected to allow free hand movement for gesturing during conversation.

The corpus was recorded over three days with between two and four participants from a pool of five on camera at any time.

3.6.3 TableTalk Corpus

The TableTalk corpus contains multimodal recordings of free flowing natural conversations among five participants, recorded at the Advanced Telecommunication Research Labs in Japan in 2007 (Campbell, 2009). In order to collect as natural data as

possible, neither topics of discussion nor activities were restricted in advance. Three sessions were recorded over three consecutive days in an informal setting over coffee, by three female (Australian, Finnish, and Japanese) and two male (Belgian and British) participants (Jokinen, 2009). All participants were native or near-native speakers of English and the conversations were held in English, although there were a few occasions where Japanese words were used during the discussion of life in Japan and Japanese culture.

Name	Sex	Age
P1	M	60's
P2	M	50's
P3	M	30's
P4	F	30's
P5	F	40's

Table 3.4 – TableTalk Participants

Table 3.4 shows the participant demographics. P1, P2, and P3 knew one another as did P4 and P5. P1 also appeared in DANS and D64. Figure 3.3 shows the participants in one of the sessions.

3.7 Preparation of Data and Synchronization of Recordings

All three corpora were prepared for annotation in the same way. The various recordings needed synchronization to determine an accurate timeline for the corpus, and to ensure alignment of video and audio data. For each corpus, there were several audio and video recordings on different cameras, many of which had different capacity for data storage. For example, recordings from Camera 4 from the D64 sessions were over two hours long, while Camera 3 in the same session recorded blocks of 25 minutes. The audio recordings were mostly created using a multitrack system and thus were already synchronised with one another, but needed to be synchronised with the video recordings. Three recordings were watched for each session and notes taken of easily

Figure 3.3 – Still from composite video showing face-recognition on participants in 5-party TableTalk conversation.

identified actions and the utterances around them. These time points were used to set a global time for all recordings in each corpus and individual recordings were then synchronized to this global time. The first step was to establish clear salient points to 'anchor' times. The timescale of events of interest to speech researchers can be quite small - a syllable is regarded as a 'long-time-span' feature and is in the order of 120 - 250 ms, average short vowel length is around 70 ms and study of prosodic variations requires granularity in the millisecond range. It was therefore vital that these anchor points be identified as exactly as possible. As audio data can be visually examined as a waveform or spectrogram, all audio tracks were extracted from video recordings using Audacity software (Audacity, 2014) and the audio tracks were used to synchronize the recordings. Points of interest were identified at phonetically salient places on spectrograms of audio tracks generated using Praat software (Boersma and Weenink, 2010). The local times for these points were measured and offsets between the various recordings determined. The audio track from video and the audio tracks from the microphones were used to manually match up audio waveforms, thus providing a synchronisation between the multiple recordings for each timepoint in each of the corpora. After the initial synchronisation, recordings were cut into 15 minute segments with 30 second buffers and these shorter recordings were then synchronized again, in order to minimise any drift between different recordings. These offsets were then checked on the video recordings using Elan software (Wittenburg et al., 2006). The transcription and annotation processes used on the data are described in Chapters 4 and 7.

3.8 The TEAMS Corpus

This section describes the TEAMS Corpus (Litman et al., 2016), used to contrast casual and task-based talk in Chapter 9.

The TEAMS corpus was created at the University of Pittsburgh and comprises 47

hours of multiparty interaction from 62 teams (35 three-person and 27 four-person), playing a collaborative board game – Forbidden Island. This game requires cooperation and communication among the players to win as a group. 213 native speakers of American English (79 males and 134 females) aged 18 years or older participated in the recordings. Before each recording session, participants took a pre-game survey collecting personal information. Each game was played for a maximum of 35 minutes, and all teams played the game twice (Game 1 and Game 2). After each game, the participants took a post-game survey to evaluate the team process.

In the game, players take on the roles of adventurers seeking treasures on an island before it is flooded. Each player takes on a role in the team, for three-party games - Pilot, Messenger, and Engineer, with a fourth role, Explorer, in four party games. The corpus was originally recorded to study entrainment in multiparty dialogues; the game was chosen as it requires players to communicate to achieve their goals, thus lending itself directly to eliciting entrainment. As each player took on a different role, the corpus creators claim that the communicative situation was very similar to real world teamwork. The game could equally well be played with three or four players, and the three party version was played when a fourth participant did not show up for recordings. The difficulty of the game was adapted to novices.

The games were played in two conditions, Control and Intervention. The Intervention condition differed in that players were given ten minutes of training in cooperation strategy and were allowed extra time between Game 1 and Game 2 to discuss how their performance could be improved.

The audio recordings were segmented into interpausal units (IPUs) and transcribed manually using the Higgins Annotation Tools (Skantze, 2009).

3.9 Conclusions

This chapter has described the background to corpus collection for analysis of conversation, outlining issues related to ecological validity, recording conditions and equipment, and multimodality. The collection of conversations comprising the dataset for analysis in following chapters has been described, and initial preparation and segmentation has been reported. The TEAMS Corpus has also been briefly described.

Chapter 4

Annotations and Dataset Preparation

"I often say that when you can measure what you are speaking about, and express it in numbers, you know something about it; but when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meagre and unsatisfactory kind; it may be the beginning of knowledge, but you have scarcely, in your thoughts, advanced to the stage of science, whatever the matter may be."

- Kelvin, 1883

4.1 Introduction

In this chapter, issues relevant to the annotation of conversation are reviewed, and the segmentation, transcription, and annotation of chat and chunk phases, laughter, and disfluency in the CasualTalk data are described, as are the annotations of floor state in both the CasualTalk and TEAMS datasets. All Python and Praat scripts used are available at `https://github.com/egilmartin/casualtalk.git`.

Conversation	Corpus	Participants	Duration (seconds)
A	D64	5	4164
B	DANS	3	4672
C	DANS	4	4378
D	DANS	3	3004
E	TableTalk	4	2072
F	TableTalk	5	4740

Table 4.1 – CasualTalk dataset for experiments

4.2 Overview of final CasualTalk Dataset

To explore the structure and characteristics of multiparty casual conversation, six conversations were drawn from three corpora of unconstrained social or casual conversation. The total length of the conversations was approximately 6.5 hours. There were two conversations each with 5, 4 and 3 participants. Individual conversation length averaged just over one hour, with range from 34 minutes to 77 minutes. Details of the final dataset are shown in Table 4.1

4.3 Considerations in Segmentation and Annotation of Spoken Dialogue

In order to analyse a corpus of speech, the data must be put into usable form, synchronized, segmented, and annotated according to the purpose of research. Annotations of spoken corpora involve segmentation of the audio stream into stretches of speech and silence, from which transcriptions of speech and annotations of utterance, turn, temporal features such as the length and position of pauses, gaps and overlaps, and labels for phenomena such as laughter or breathing can be generated.

The use of existing annotations is very attractive to researchers as annotation is costly and extremely time-consuming. However, most annotations are done for a particular purpose and therefore may not catch the phenomena of interest to later

studies. For the current work, accurate segmentation was needed, in addition to transcription of speech and labelling of laughter, disfluency, and chat/chunk phases.

4.3.1 Unit of Segmentation

The first step in this process is segmentation. Here the audio is marked up into intervals of speech and silence. A major element in segmentation and indeed in all analysis of temporal properties of conversation is the basic unit of speech and silence chosen (Heldner et al., 2008). The inter-pausal unit (IPU) is widely used, particularly for automatic segmentation, and defined as a stretch of talk or talkspurt by a speaker bound on either side by silence from that speaker. A threshold is set for the length of within speaker silence required for the definition, generally at around 80-120ms in order to eliminate 'false positives' due to silences generated by stop occlusions. In this study, segmentation was manual, and thus within speaker pauses due to stop occlusions could be marked as such and discarded from the data as necessary. Importantly, it was possible to perform this segmentation at the finer intonational phrase (IP) level, where each segment covered one intonational phrase or prosodic 'tune' from a speaker. This level of segmentation is valuable in analysis of turntaking as it allows finer analysis of *when* a second speaker chooses to speak in backchannel or interruption of a speaker holding the floor. Intonational phrases can easily be concatenated to generate an IPU segmentation, but when segmentation is carried out only at IPU, utterance, or turn level, any finer detail is lost and the data would need re-segmentation to identify intonational phrases.

4.3.2 Manual or Automatic Segmentation and Transcription?

Manual segmentation is a very time-consuming and therefore expensive process, particularly for large amounts of data. Therefore, automatic segmentation using Voice Activity Detection (VAD) or Speech Activity Detection (SAD) is often used for segmen-

tation. Many annotation software packages include facilities to perform automatic segmentation to speech/silence (e.g. Praat, ELAN), or voice/speech activity detection (VAD/SAD) software can be used for the same purpose. It is important to note that many of these applications (VAD) do not in fact separate the audio into speech and silence but rather to sounding and silent intervals, where sounding refers to periods where there is strong harmonic activity in the speech wave - that is during vowel production or production of voiced consonants. A significant proportion of speech is unvoiced and therefore marked as silent by the software - thus utterance boundaries produced are not exact and may be misleading in cases where there is unvoiced speech such as aspiration noise or a fricative at the beginning or end of an utterance. More sophisticated SAD algorithms can distinguish voiced/unvoiced sounds, but only work well with clear audio, which is not the case with the data used in this thesis.

Any automatic or manual annotation is an abstraction and it is vital to know how this abstraction will affect the phenomena of interest - for example in studies based on utterance length, a shortening of utterance length in all cases where the utterance begins and ends in unvoiced phonemes could cause major inaccuracies and introduce artefacts to the data if there were a significant number of such utterances in the data. It is also important to note that automatic segmentation produces at best an IPU-level segmentation, rather than a finer IP-level segmentation.

Another issue in any form of dialogue is that any automatic annotation depends how 'clean' the recording is. When speakers in proximity are being recorded, speech from other participants can be picked up by a speaker's microphone, resulting in 'bleedover' or crosstalk appearing in the recording. Ideally what is needed is a recording of each speaker made using a close microphone (head-mounted if possible) with enough separation to result in an audio file which contains little or no bleedover or crosstalk. Recordings of this nature can be obtained by recording conversations

where speakers are in separate booths, or speaking on the telephone. However, the resulting recordings are not of natural face-to-face interaction. If recordings are made in face-to-face situations using close coupled microphones and do contain bleedover, post-processing can sometimes be applied to clean up the signals. The basic principle underlying such speaker diarization methods is that the speech of interest will be significantly louder or the signal will contain more energy, essentially using the same technology as used in speech/silence detection.

With clean recordings - one speaker recorded in a quiet environment - such as those used in fine-grained phonetic research, automatic methods can be used quite successfully for segmentation to speech/silent intervals, and are often included in annotation packages - for example, this is achieved in Praat by setting a threshold level of noise as the maximum for silent intervals. Even with recordings made in quiet but not completely silent environments, a trade-off has to be made as much of the speech signal consists of high frequency unvoiced sounds (plosives, fricatives, etc) which do not carry nearly as much amplitude or power as voiced sounds - these are often lost in automatic segmentation and replaced by manual checking of the data. With multiparty recordings made in naturalistic surroundings automatic segmentation is much less accurate.

The amount of overlap in the data also makes acoustic analysis difficult. An intermediate strategy is sometimes used whereby clean stretches of data from individual speakers are used for acoustic analysis. However, many of the phenomena of interest occur just in the 'dirty' stretches, where there are turn changes or overlapping speech, just where automatic post-processing methods are least effective, and thus extreme care must be taken in analyses of these types that the clean data used for analysis is actually representative of the phenomena of interest. In addition, in analysis of conversation, the phenomena of interest may be backchannels where the speaker's output may be lower than that of other speakers output bleeding over onto the recording. In

light of these issues, for the current study it was decided to manually segment speech and silence in the dataset, a process which took several months. A comparative study of how automatic methods perform versus manual segmentation on a subset of the data was carried out as reported in Section 4.4.2.

The efficacy of machine segmentation was found to be completely inadequate for the data studied, particularly as, the locations where error is likely to occur are precisely around floor changes, the very phenomena of interest to the thesis. While machines can function quite well on stretches of single party speech, they fail where there is a lot of overlap between speakers. So, simply running a VAD/SAD over the data may give an unfairly positive view of the efficiency of the process as the system would be largely correct for single party speech in the clear and for silence, which form the bulk of the conversations, but would fail around overlap, an area of interest.

The results of tests with Opensmile , described in Section 4.4.2 showed approximately 50% error rate on speech silence discrimination(segmentation) around floor state changes on relevant data. As the focus of the work is using speech and silence statistics to better understand multiparty conversation, a 50% is unacceptable. Machine error rates would be lower for two party speech recorded in clean (but less natural) conditions. As segmentation was manual, manual transcription was carried out at the same time, and thus automatic transcription was not needed

Many of the issues with the data described above with respect to segmentation also apply to transcription. For very large datasets, manual transcription is extremely expensive, and there are advantages to the consistency of automatic transcription which may offset the inaccuracies introduced by its use (Ernestus and Baayen, 2011), but in this case, the manual segmentation made it logical to transcribe manually at the same time, and thus the transcriptions were performed by human annotators.

4.3.3 Annotation Tools

To annotate video and audio files, various software applications are available. For audio annotation Praat (Boersma and Weenink 2010) is well known and widely used in the linguistic community, and was used in the present work for audio annotation. Tools considered for annotation of video were Anvil(Kipp, 2001), Elan ((Wittenburg et al., 2006), and Vcode (Hagedorn et al., 2008).

Annotating video and audio are similar in many ways, but differ in one important characteristic - in audio annotation the progression of the signal can be seen over a time span as a static image by using the visual dimension – the displayed waveform and spectogram and other acoustic detail (pitch curves, etc) of the sound being annotated. Thus, much of the fine annotation work is not done by ear alone but rather with reference to the spectrogram and waveform visible onscreen and by triangulation of this information with the heard signal. In video annotation, there is no abstracted representation of the signal such as a waveform or spectrogram and thus annotation must proceed by first watching and then pausing, and watching frame by frame to annotate. Anvil deals with this problem by providing a 'storyboard' of frames of the video, allowing for faster annotation. In systems without this storyboard, annotation can be very time-consuming, but can be sped up by annotating 'on the fly' on very slow playback. This relies on a segmentation functionality that allows keystrokes to be assigned to annotation labels so tiers can be marked as they play, and also permits the user to set a lag to compensate for delay caused by human reaction time. This functionality is available in Elan and Vcode, but Vcode is only available for MacOS while Elan is multi-platform. Anvil was found to be difficult to install due to issues with codecs (media encoder/decoder software) on different operating systems, while Elan installed without problems on all computer systems tried. Elan is also popular in the dialogue community and annotations made in Elan can easily be shared and used in collaboration with other researchers.

Elan was therefore chosen as the video annotation tool for the current work, while Praat was used for audio annotation.

4.4 Segmentation Procedure

Segmentation into speech and silence on each speaker track is a fundamental step in data preparation. 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, in addition to noise introduced due to the naturalistic setting of the conversations, such as kettles boiling, cups clattering, or traffic noise from open windows. While automatic segmentation could handle stretches where only one or two participants were talking without overlap, many of the turn changes involved overlap and there was significant choral production of short utterances and laughter. While there has been much progress in the automatic analysis of spoken conversation (Heldner and Edlund, 2010; Laskowski, 2011), attempts to automatically segment the dataset in this study were not successful. Although rough automatic segmentations were produced at the segmentation stage using Praat, annotators found it faster to segment manually than to correct boundaries on this automatically segmented data.

4.4.1 Manual Segmentation

For segmentation and annotations purposes the sound files were cut into fifteen minute WAV segments to lower the computational load and to allow applications to work faster. Praat annotation software (Boersma and Weenink, 2010) was used to create an annotation file or *Textgrid* for each 15-minute segment, containing annotation tiers for each of the speakers and an 'unknown' tier for doubtful cases where it was unclear on the audio who was speaking - for example, when a speaker's microphone picked up some of a different speaker's speech. The audio files were then segmented manually

into speech and silence intervals on the relevant TextGrids. The segmentation was carried out at the intonational phrase level (IP), rather than a more coarse utterance or inter-pausal unit (IPU) level. The Praat textgrids contained one tier per speaker, and each speaker's production was segmented separately. After a preliminary trial segmentation of part of the data, a final label set was devised to cover all participant contributions. The labels covered speech (SP), silence (SL), coughs (CG), breaths (BR), and laughter (LG). The speech label was applied to verbal and non-verbal vocal sounds (except laughter) and thus included contributions such as filled pauses such as 'um' or 'uh', short utterances such as 'oh' or 'mmhmm', and sighs. Laughter was annotated inline with speech.

The annotator listened to 10 second and, where necessary, four- or even two-second windows of each sound file on a wav and marked on the relevant tiers the intervals in which the main speaker was speaking and any audible speech by other speakers. Where the speaker could not be identified the interval was marked in the unknown tier. 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 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 and conclusively identify who was speaking.

While the data under consideration make automatic detection of silence impossible, there are valid concerns about manual segmentation into speech and silence. It has long been known that humans listening to speech can miss or indeed 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 these early results were based on speakers timing pauses with a stopwatch in a sin-

gle hearing. In the current work, using Praat and Elan, speech can be slowed down and replayed and, by using the four-second window, annotators can see silences or more accurately differences in amplitude on the speech waveform and spectrogram. This option 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. Therefore, it was much more likely that silences would be picked correctly and it was hoped that problems due to annotators only picking up perceptually salient silences could be avoided. A simple measurement test was performed to check the precision of Praat software, by physically measuring the minimum step possible in moving boundaries, on fullscreen display on a 23" monitor, which was the screen size used for segmentation. On 10 seconds of audio, there was a possible error of 50 milliseconds, and 20 milliseconds for 4 seconds.

The textgrids resulting from the segmentation were concatenated into one per conversation. A Praat script was written to concatenate contiguous identical labels, thus creating a second segmentation into approximate inter-pausal units. Although breath is extremely interesting as a feature of conversation, particularly as a cue for turntaking, it was not possible to annotate breath accurately for all participants and thus the breath intervals annotated were converted to silence for the purposes of this study. Similarly, coughs were relabelled as silence for the current work.

4.4.2 Automatic Segmentation for Comparison

Analysis with the OpenSMILE toolkit[1] (Eyben et al., 2013a) implementing the SAD described in (Eyben et al., 2013b) confirmed that even under artificially good conditions the best average error per segment was around half the duration of the silences to be studied. [2] These results found using a state of the art system confirmed that conver-

[1] SMILExtract 2.3 with vad_opensource.conf, post-processed with a standard cRnnVad2 component configuration

[2] False alarms were filtered under the premise that a beam-former would be able to decide whether speech activity came from the targeted speaker or from cross talk. Short gaps under 60ms were bridged

sation of this type still needs at least manual correction of segmentations. The Speech Activity Detection was carried out with the help of Christian Saam of the ADAPT Centre, Trinity College Dublin.

4.5 Annotation of Floor State

To allow analysis of floor state changes and overlapping speech and laughter in the dataset, a floor state tier was then created for each conversation, using a custom Praat script adapted by the author from an existing script by Jose Joaquin Atria. This tier divided the entire conversation into labelled intervals, for each interval recorded who was speaking or laughing during the interval timespan, and also labelled intervals of global silence. For the floor state intervals, an alphanumeric code was assigned to show which participants were active and what they were doing (speaking or laughing) - this step was taken so that the resulting strings could be manipulated using regular expressions (regex) to extract relevant information for analyses on overlap and speaker change patterns. For example, the label **aSPbSPcLG** denotes that speakers **a** and **b** are speaking and **c** is laughing for the duration of the interval. In cases where there was nobody speaking or laughing, the label was **GS** for global silence.

Figure 4.1 shows an example of this tier. This floor state annotation was also carried out on the TEAMS corpus.

4.6 Transcription

The conversations were also fully transcribed, and this transcription tier was used for automatic alignment of the audio with the transcription, and to annotate disflu-

and a collar of 30ms was applied around the segment boundaries during scoring. The speech activity detection was scored with NIST SCTK 2.4.0 md-eval.pl version 22 according to the guidelines of the Spring 2006 (RT-06S) Rich Transcription Meeting Recognition Evaluation Plan, Section 7 DIARIZATION — 'SPEECH ACTIVITY DETECTION' MDE (Fiscus et al., 2006) (n=3142 missed=279.34s falarm=1753.06s total=2032.4s avg. missed=0.0889052s/1 avg. falarm=0.557944s/1 avg. total=0.646849s/1)

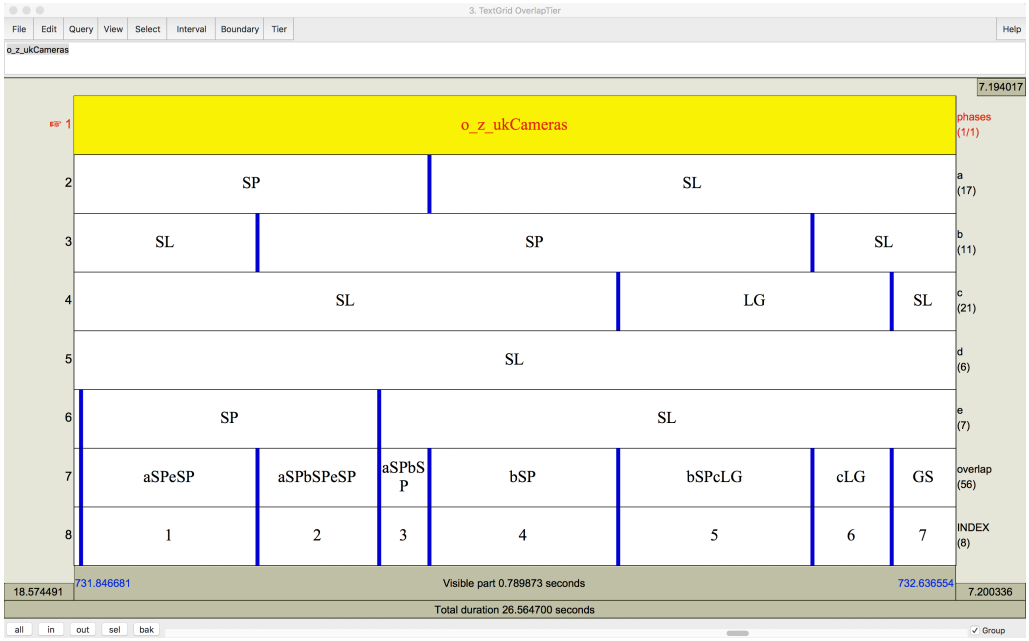


Figure 4.1 – Extract from Praat textgrids for Conversation A, showing the floor state tier

encies, The transcription was carried out at the IP level, and all words, incomplete words, and vocal sounds were marked. As with the segmentation, the use of automatic transcription was found to be inefficient for this particular data set.

4.7 Annotation of Chat and Chunk Phases

Chat and chunk phases were marked using an annotation scheme devised from the definitions of chat and chunk phases given in Slade and Eggins' work (Eggins and Slade, 2004), and the descriptions given by Slade in her PhD work of transcripts of chat and chunk phases in casual talk (Slade, 2007).

For an initial classification, conversations were divided by first identifying the 'chunks' and considering everything else chat. In the first instance, this was done using the first, structural, part of Slade and Eggins' definition of a chunk as "a segment where one speaker takes the floor and is allowed to dominate the conversation for

an extended period." (Eggins and Slade, 2004). These chunks could then be further classified into their respective genres.

4.7.1 Annotation Guidelines

The following guidelines were created to aid in the placing of chat/chunk boundaries.

Start A chunk starts when a speaker has established himself as leading the chunk. So, in a story, the speaker will be starting the story, and the story will be seen to go on. Of course, this involves 'telling the future' and couldn't be done by a machine.

Stop To avoid orphaned sections, a chunk is ended at the moment the next element (chunk or chat) starts.

Aborted In cases where a story is attempted, but aborted before it is established, this is left as chat. In cases where there is a diversion to another element mid-story, for example, and a return later, all three elements are annotated as though they were single chunks/stretches of chat but the chunks are given the same title with indices to distinguish them.

Overlap Sometimes a new chunk begins where a previous chunk is still tailing off. In these cases the new chunk onset is the marker of interest and the old chunk is finished at the onset of the new one.

Once the chunk was identified, it could be classified by genre. This process was aided by the taxonomy of the stages of various types of story, given by Eggins and Slade (Eggins and Slade, 2004) as shown Table 4.2.

4.7.2 Annotation Scheme

For annotation, a set of codes for the various types of chunk and chat was created. Each code is a hyphen-separated string containing at least a Type signifier for chat or

Genre	Generic Structure
Narrative	(Abstract) ^ (Orientation) ^ Complication ^ Evaluation ^ Resolution ^ (Coda)
Anecdote	(Abstract) ^ (Orientation) ^ Remarkable Event ^ Reaction ^ (Coda)
Exemplum	(Abstract) ^ (Orientation) ^ Incident ^ Interpretation ^ (Coda)
Recount	(Abstract) ^ Orientation ^ Record of Events ^ (Coda)
Observation/Comment	(Orientation) ^ Observation ^ Comment ^ (Coda) ^ (Completion)
Opinion	Opinion ^ Reaction ^ (Evidence) ^ (Resolution)
Gossip	Third Person Focus ^ Substantiating Behaviour ^ (Probe) ^ Perjorative Evaluation... ... ^ (Defence) ^ (Response to Defence) ^ (Concession) ^ (Wrap-up)

Table 4.2 – Generic structure for 'chunk' types amenable to such analysis, () indicate optional stages

Generic labels	
Type	Chunk: x, Chat: o
Ownership	mark with corresponding speaker code from transcript - a, b,... for chat - z
Story	S Can be further marked with codes denoting subtypes: Narrative (N), Anecdote (A), Exemplum (E), Recount (R)
Observation/Comment	C
Opinion	O
Gossip	G

Table 4.3 – Labelling scheme for chunks

chunk, an Ownership label, and optional subelements further classifying the chunks as shown in Table 4.3.

4.7.3 Annotation Procedure and Inter-rater Agreement

The chunks were first roughly identified from the transcriptions. A Python script was written by the author and used in conjunction with a Praat script by Miette Liennes adapted by the author to generate a Conversational Analysis style transcription from a Praat textgrid with different coloured text for each participant in the conversation. The resulting transcriptions were read and the onset and offset of chunks marked off. The type of chunk was then decided with reference to Slade and Eggins' taxonomy. The conversations were then listened to in Praat and the temporal boundaries of each chunk marked off on a 'phases' tier. Intervals were labelled using the code shown in

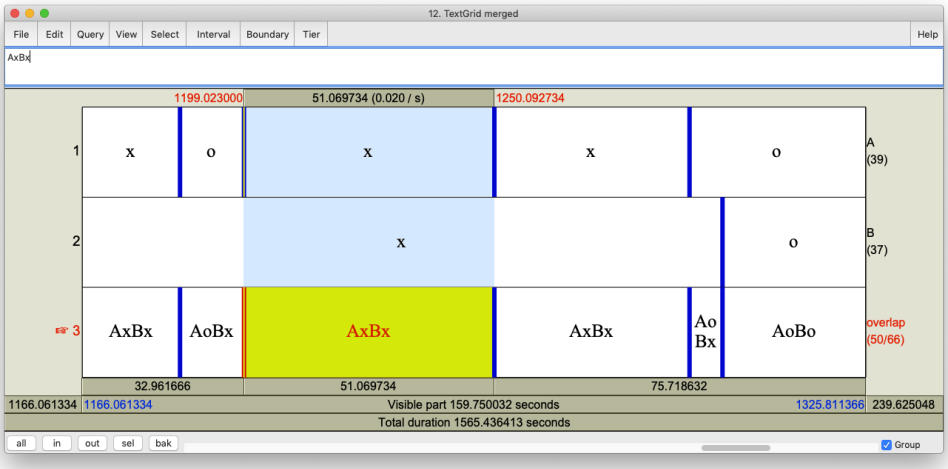


Figure 4.2 – Praat Textgrid showing inter-rater labels for chant and chunk annotations. The A and B tiers are combined to create 'overlap' labels which record A and B's annotation for the interval covered.

Table 4.2, marking type of phase (chat - o or chunk - x), subtype of phase (narrative, story, discussion..), name of phase (roughly equivalent to the topic under discussion), and phase 'owner' (main speaker in chunks and everyone in chat phases). As an example, the code **x_s_g_cats** would denote a chunk phase where the main speaker is **g** and the chunk, which was in story form, was about cats. It was found that chunks could generally be identified quite well from reading the transcripts.

A 25-minute section of the corpus was annotated for chat and chunk by two annotators, in order to assess the robustness of the annotation scheme. Inter-rater reliability was obtained using Cohen's Kappa using the timespans marked as chat or chunk. As segmentation into phases and labelling were carried out by annotators in one pass, inter-rater reliability could not be assessed by simply matching these labels, as the start and end points of each annotator's labels were chosen by the annotator themself during the joint segmentation and labelling process. To address this problem, a new label set was generated from the annotator labels in order to use time as a proxy for label counts. The Praat script for floor state change mentioned in Sec-

		A	
		x	o
B	x	876.55	136.73
	o	57.96	494.20

Table 4.4 – Confusion Matrix for Chat (o) and Chunk (x) label duration

tion 4.5 was adapted to create a set of labels from the annotation tiers for chat and chunk phases, as shown in Figure 4.2. The labels reflected how each annotator had marked the relevant segment – there were four 'base' labels Ax, Bo, Ax and Bx where A or B denoted the annotator and o or x denoted chat or chunk phases, which were then combined into 'inter-rater' labels showing who had marked what for each segment - for example a segment marked AxBx was considered chunk by both A and B, while an AxBo label denoted a segment which A considered chunk and B considered chat. These resulting 'overlap' labels - AxBx, AxBo, AoBx and AoBo were then used to calculate agreement between the raters. To do this, the duration of each label was essentially used as a count of a number of contiguous shorter labels (of .01 second in duration). While there are reservations in the literature about the strength of the independence assumption on contiguous labels in the calculation of Cohen's Kappa for speech corpora (Rietveld et al., 2004), in this case the short derived labels were not individually chosen by the annotators, but rather a division of the much larger chat and chunk labels annotated.

The inter-rater label durations were summed to generate the confusion matrix in Table 4.4. Agreement between A and B (the sum of label durations where A and B agreed divided by the total duration) was 87.56%. Cohen's Kappa was calculated from the confusion matrix, using the formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{4.1}$$

where:

Po is the observed proportional agreement (the total duration of AoBo and AxBx labels divided by the total duration of all labels),

Pe is the hypothetical probability of chance agreement: the sum of the probability of both raters randomly choosing x (the total duration where A chose x multiplied by the total duration where B chose x)

Ax = total time A chose chunk = $\sum AxBx, AxBo$

Bx = total time B chose chunk = $\sum AxBx, AoBx$

Pchunk = probability that both raters would randomly choose chunk

$$= \frac{Ax}{totalduration} * \frac{Bx}{totalduration} \tag{4.2}$$

Ao = total time A chose chat = $\sum AoBo, AoBx$

Bo = total time B chose chat = $\sum AoBo, AxBo$

Pchat = probability that both raters would randomly choose chat

$$= \frac{Ao}{totalduration} * \frac{Bo}{totalduration}$$

Pe = Pchunk + Pchat

The resulting Kappa was 0.74, which fell into the 'substantial' band (0.61–0.80) of the interpretation scale proposed in (Landis and Koch, 1977) based on Cohen's original work (Cohen, 1960), and the moderate (.60–.79) band of the more stringent scale proposed in (McHugh, 2012) for medical labelling. In conjunction with the raw agreement score of 87.56%, this was deemed acceptable as evidence that there was moderate to strong agreement between the annotators.

A total of 213 chat and 358 chunk phases were identified across the six conversations.

4.8 Annotation of Disfluency in Conversation A

Conversation A from the dataset was chosen for analysis of disfluency as the signal quality was sufficient for automatic forced alignment. Forced alignment of the data provides approximate timestamps at the phoneme and word levels and thus allows analysis of temporal characteristics of disfluencies and also of the distribution of disfluency in chat and chunk phases. The conversation was force aligned to approximately segment the data into phonemes and words, by running the Penn Aligner (Yuan and Liberman, 2008) over a sound file and accompanying transcription for each intonation phrase annotated. The data were then annotated for disfluencies using Shriberg's Pattern Labelling System (Shriberg, 1994, 53), as described in Section 2.6.2 of Chapter 2.

4.8.1 Preparation for Annotation – Forced Alignment

Automatic alignment was performed with Penn Phonetic Labs Forced Aligner Toolkit (P2FA) in conjunction with the Hidden Markov Model Toolkit (HTK) version 3.4 (Young et al., 2006) in order to obtain an estimate of where word and phoneme boundaries fell within each utterance.

The Penn Phonetic Labs Forced Aligner Toolkit (henceforth referred to as the P2FA) is a freely available open source forced phonetic aligner developed at Penn University Phonetics Labs (Yuan and Liberman, 2008). The toolkit is based on HTK for word/-phoneme recognition using the CMU Pronouncing Dictionary, an open-source machine readable pronunciation dictionary for American English (The CMU Pronouncing Dictionary, 2007). The main component of the toolkit is a Python script which takes a textfile with orthographic transcription and the corresponding WAV audio file as input. The script uses HTK with the pronunciation dictionary to generate a Praat compatible TextGrid file with two tiers containing time-aligned phones and words from the utterance.

4.8.2 Text Normalisation for Alignment

The P2FA requires text to be uppercase and in US spelling. As it is based on speech recognition, it can only recognise words in its lexicon - in this case the CMU dictionary. Therefore, text normalisation and resolution of out of vocabulary words were necessary before the transcriptions could be passed to the aligner. Custom Praat and Python scripts were created to normalise the transcriptions by converting all text to upper case and changing any UK spellings to US spelling. After this process, a script was written to compare text files of the transcripts with the textfile containing the CMU dictionary used in the aligner, and to generate a list of out of vocabulary words and other unrecognised tokens. The list contained several words which were not in the dictionary, a number of mispellings, and all of the words marked unfinished in the transcriptions. The mispellings were corrected. Any out-of-vocabulary words were manually transcribed to the ARPABET-based phoneme set used in the CMU dictionary. These new transcriptions were added to the dictionary for use with the aligner. Fifty-nine words and transcriptions were added to the dictionary for use with the aligner - these are listed in Appendix A

The conversations contained a number of incomplete words, for example if a speaker started saying 'industrial' and only produced 'indust...' before abandoning the word, it was marked **indust-** with the dash noting that the word was incomplete. Such labels would not be recognised by the aligner, and thus all unfinished words were marked as **NS** or noise for alignment purposes. In this way they would be ignored by the aligner while the rest of the phrase was aligned. These labels were later manually transcribed and their boundaries checked on the TextGrids resulting from the aligner.

Tokens for optional short pauses, 'sp', were added between words and at the start and end of each utterance. These allow the aligner to add short pauses in the resulting transcriptions if found in the speech data. The data were then divided into into-

national phrases (IPs) and a sound file generated for each IP from the corpus recordings. The aligner was then run over the sound file and accompanying transcription for each intonational phrase in the conversation using a custom Python batch script. Sections where there was significant overlap were marked as noise, run through the aligner to create a 'dummy' TextGrid for concatenation purposes and later manually aligned. The resulting textgrids were then concatenated using a script adapted from a Praat script by Daniel Hirst, and the final textgrid was used for manual annotation of disfluencies.

4.8.3 Annotation using Pattern Labelling System

Disfluency annotation was performed using Shriberg's Pattern Labelling System (Shriberg, 1994, 53), with the following adjustments to suit the data.

1. Filled pauses were marked as **f** when they occurred within or adjacent to an utterance by a speaker, i.e. as part of a speaker's turn or attempt to take a turn. Isolated 'uh' and 'um' filled pause like productions were marked **v**, but were not considered disfluencies for this study.

2. Filled pauses appearing adjacent to (before or after) or within more complex disfluencies were included in the label for that disfluency for ease of analysis. As an example, a filled pause directly before a repetition was marked as **[fr.r]** rather than as two sequential disfluencies, **[.f][r.r]**

3. Pauses were marked as **x** if they occurred within a sentence, while isolated *uh* and *um* sounds were annotated with **v**. These productions were generally backchannels, and were annotated to allow contrasting of isolated uh and um sounds with filled pauses in later work.

4. Pauses occurring adjacent to (before or after) or within more complex disfluencies were included in the label for that disfluency, as described for filled pauses

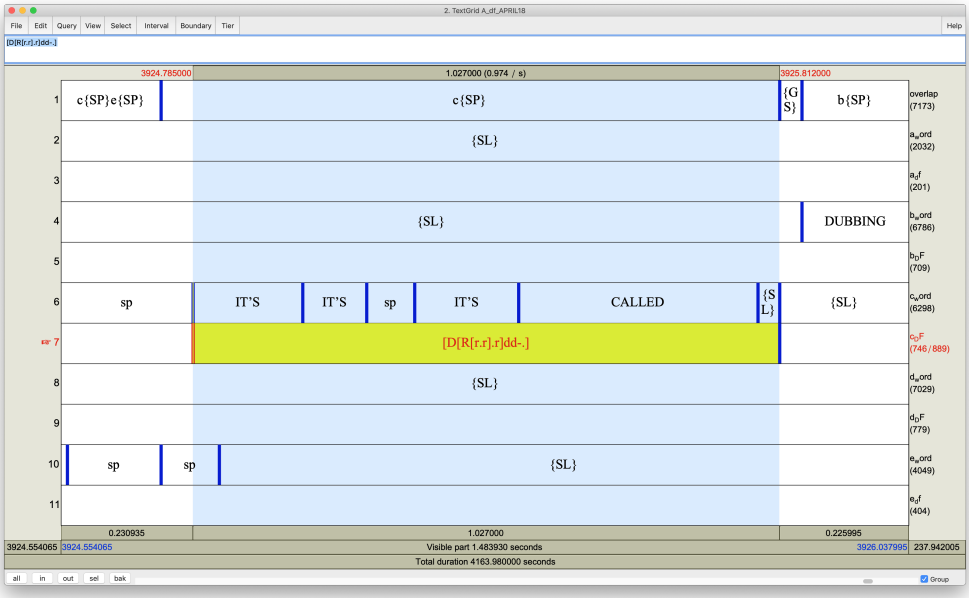


Figure 4.3 – Praat TextGrid used for annotation of disfluencies

above.

The annotation was carried out manually, for each of the five speakers in the conversation, by adding a disfluency tier for each speaker to the TextGrids generated from the aligner, as shown in Figure 4.3. The resulting disfluencies were then classified using Shriberg's Type Classification Algorithm as described below.

4.8.4 Grouping of Disfluencies with Type Classification Algorithm

Shriberg's Type Classification Algorithm (TCA) provides a method to group annotated disfluencies (Shriberg, 1994, 75). The algorithm gives an order of preference for classification of disfluencies containing more than one label. The algorithm sorts disfluencies by membership in the following order:

1. ART (articulation)

2. HYB (hybrid)

3. SUB (substitution), INS (insertion) and DEL (deletion)

4. REP (repetition)

5. CON (conjunction)

6. FP (filled pause)

The classification is motivated by the amount of change each of the classes entails, with substitutions, deletions and insertions involving changes in the wording, while repetitions do not, and filled pauses are extra syntactic. Shriberg's rules for the algorithm are shown in Fig 4.4. The TCA was applied to the disfluencies annotated in Conversation A, and the annotations were then used in the analysis of disfluency in conversation A, as described in Chapter 7.

4.9 Annotation of Laughter

Laughter was annotated in line with speech and silence. When a participant laughed the beginning and end of their laughter was marked. Annotators only marked 'full' laughter, and left speech laughs, where participants' speech was coloured by laughter with shaky vowels, labelled as speech. More detailed studies of laughter in speech would require a separate transcription tier for laughter parallel to speech tiers, thus allowing for annotations of speech laughs as simultaneous speech and laughter.

4.10 Overview of Annotations

Table 4.5 shows annotations completed for each of the conversations in the dataset.

4.11 Conclusions

The final segmented CasualTalk dataset of six conversations yielded intervals of participant contributions where each interval was a concatenation of contiguous IPs with

	Corpus	Segmentation	Overlap	Phase	Transcription	Disfluency
A	D64	Y	Y	Y	Y	Y
B	DANS	Y	Y	Y	Y	-
C	DANS	Y	Y	Y	Y	-
D	DANS	Y	Y	Y	Y	-
E	TableTalk	Y	Y	Y	Y	-
F	TableTalk	Y	Y	Y	Y	-

Table 4.5 – The annotations completed for each of the conversations in dataset

the same label, totalling 5 hours and 26 minutes of speech and 23 minutes of laughter. There were a total of 571 chat/chunk phases across the dataset, comprising 213 chat segments and 358 chunk segments. The CasualTalk and TEAMS corpora were annotated with floor state labels denoting the situation at any moment of the conversation in terms of speech and silence for each participant. The disfluency annotation of Conversation A resulted in 1586 marked disfluencies. This chapter has described the issues involved in and procedures followed for the organisation, segmentation and annotation of the data for use in the experiments reported in the following chapters.

TYPE	Must Include	Must Not Include	May Include	Examples
ART	~		s i d r c f	shle she liked it [r~.r] i'd like to fry uh to fly from boston [rr~.frr] show grand trou- ground transport [r~r~-.rr]
HYB	2^+ from: s i d	~	r c f	(example of s + i): she liked he really liked it [sr.sir]
SUB	s	~ i d	r c f	she he liked it [s.s] she uh he liked it [s.fs] and she liked and he liked it [csr.csr]
INS	i	~ s d	r c f	she liked really liked it [r.ir] she liked uh really liked it [r.fir] and she liked and she really liked it [crr.crir]
DEL	d	~ s i	r c f	it was she liked it [dd.] it was um she liked it [dd.f] and it was and she liked it [cdd.c]
REP	r	~ s i d	c f	she she liked it [r.r] she liked uh she liked it [rr.frr] and she and she liked it [cr.cr]
CON	c	~ s i d r	f	she saw it and and she liked it [c.c] she saw it and uh and she liked it [c.fc]
FP	f	~ s i d r c		um she liked it [.f] she uh liked it [.f]

Figure 4.4 – Shriberg's Type Classification Algorithm used in the current work to classify annotated disfluencies (Shriberg, 1994, 76). Deletions (d), filled pauses (f) which were part of a turn, hybrids (h), insertions (i), repetitions (r), substitutions (s), isolated filled pauses (v), and unfilled pauses (x)

Chapter 5

Overview of Methods

In this section, the measures and statistical methods and conventions used in the experimental work and the dataset are described.

5.1 Statistical Methods and Conventions

Much of the experimental work ahead involves comparison of distributions, and investigation of whether and how these differ. These tests are performed in order to better understand how chat and chunk phases of casual talk differ in terms of a set of basic features - speech, laughter, disfluencies and floor state dynamics or activity around silences and overlap, and how casual and task-based talk compare. Thus the distributions are of parameters including duration (of speech, overlap, laughter, silence, and chat and chunk phases), frequency of occurrence (of laughter, speech, silence), measures of liveliness (compression, floor state change rates). These parameters are contrasted for differing conditions, including co-occurrence in chat vs chunk phases, and in casual vs task based talk. The aim of the analysis for chat and chunk phases is not to prove the existence of these phases, which have been annotated for chat and chunk, but rather to see the quantitative differences in various features, information which has not been known, as chunks have been recognised using a high

level description. The standard practice in comparing such parameters is to obtain probability distributions for features of interest in each of the conditions of interest, and use statistical tests to assess whether random samples from either distribution are more or less likely to be drawn from a single distribution. When the probability that a sample is drawn from the same distribution as another is lower than a threshold level, it is possible to reject the null hypothesis that both samples belong to the same population distribution. If the sample distributions approximate known curves (Normal, Poisson, etc), then the known parameters of such curves can be used to model the distributions in question and parametric tests can be used.

In their raw state, the measures encountered in the work in this thesis are often not amenable to description using parametric statistics, and statistics based on approximation to a Normal (Gaussian) curve in particular. In many cases, particularly for duration measures, the values have a hard left boundary of zero, often resulting in distributions with heavy right skew and elevated means. In some cases, conversion of raw values to log form will result in a distribution which approximates a Normal curve well enough to allow parametric modelling. However, this log-normality is not a given. For duration data in their raw state, the mean is not a good measure of central tendency as it tends to be elevated due to the long right hand tail of large values. However, the median is less susceptible to the influence of this long tail, and often gives a better measure of central tendency. In other cases, the distributions do not approximate known distributions. In the bulk of the work ahead, non-parametric statistical tests are used on raw values to compare distributions.

Overviews of proportion, for example how conversational time is divided between silence, one-party speech and overlap, are generally given in percentages. When frequency distributions are log-transformed the natural log is used. When frequency distributions are used, whether raw or log-transformed, outliers are defined as values above 1.5 times the Interquartile range.

When comparing distributions grouped into two categories, for example, comparing durations of chat and chunk phases, parametric Welch Two Sample T-tests are used if the distributions meet the test's assumptions of near-normality and homogeneity of variance. Otherwise, non-parametric Wilcoxon Rank Sum (aka Mann-Whitney U) tests are used. These tests exist in two-tailed and one-tailed versions, where the default two-tailed test assesses whether distributions differ, while a one-tailed test tests whether one distribution is greater or smaller than another. When comparing distributions grouped into more than two categories, as an example when checking if chunk duration differs significantly by speaker (where there are 24 speakers), Kruskal-Wallis tests, a non-parametric alternative to one-way analysis of variance (ANOVA) are used when the ANOVA assumptions are not met.

The method followed in most cases is to set a null hypothesis that there is no significant difference between the sample distributions in question, and an alternate hypothesis that the distributions differ significantly. If the probability that the samples came from the same overall distribution is lower than a threshold value, the null hypothesis is rejected. For tests of correlation, Kendall's Tau, a nonparametric analogue to the Pearson Product Moment Correlation, is used.

All analyses are performed using the R statistical package (R Core Team, 2017). In the interests of reproducible research, the dataframe, Python, Praat, and R files used to prepare and analyse the data are available on Github [1].

5.2 CasualTalk Dataset

The final casual conversation dataset comprises 6 conversations drawn from three corpora of long form multiparty casual conversation, a total of 6 hours and 24 minutes of conversation. Table 5.1 provides an overview of the conversation sources, participant gender, and conversation length. Full descriptions of the corpora used, the

[1] https://github.com/egilmartin/casualtalk.git

data preparation process, and the annotations performed are given in Chapters 3 and 4

Conversation	Corpus	Participants (Gender)	Duration (Seconds)
A	D64	5 (2F/3M)	4164
B	DANS	3 (1F/2M)	4672
C	DANS	4 (1F/3M)	4378
D	DANS	3 (2F/1M)	3004
E	TableTalk	4 (2F/2M)	2069
F	TableTalk	5 (3F/2M)	4740

Table 5.1 – Dataset for experiments - conversations, participant number and gender, duration in seconds

The speech, silence, laughter, phase, and overlap information from each conversation was extracted in tabular form from Praat. These tables contained information on the Praat tier each observation came from, the start and end times for each interval observed, and the label for the interval. A column was added to each table to identify the conversation (labelled A to F) and the resulting tables were concatenated to form one large table. The transcription and disfluency labels were added to the table, and the resulting file, comprising approximately 74,000 observations, was imported to the R's programming environment as a dataframe for statistical analysis.

In R, a number of extra variables or columns were generated - Table 5.2 summarises the final dataset.

Variable	Type	Description
conv	char	Index of conversation from A to F
tmin	num	Start time for interval
tier	char	Praat tier interval came from
		a,b,.. for speakers, overlap, phases
label	char	Label of Praat interval:
		speakers: SP, SL, LG for speech, silence, laughter
		overlap: id of speakers in overlap or GS for general silence
		phases: code for chat or chunk beginning with o or x
		with 5th character in chunk labels denoting owner of chunk
tmax	num	End time for interval
tierf	factor	Tier factorized
dur	num	Duration of interval
label2	char	Same as label but has "x" or "o" only for phases
dlast	num	Distance from beginning of enclosing phase
		result of running get_nl(phases) on each conversation
dnext	num	Distance from end of enclosing chat or chunk phase
t_name	char	Name of enclosing chunk or chat phase
owner	factor	Generated using get_owner function from t_name
		gives the owner of current phase if chunk, everyone if chat
chatchunk	factor	Whether item is part of chat (o) or chunk (x) phase
labelf	factor	label2 as factor
spgen	char	Gender of speaker (for speaker tiers)
chgen	char	Gender of chunk owner
numsp	num	How many speaking in interval if overlap tier
numlg	num	How many laughing in interval if overlap tier
floor	num	Combination of numsp and numlg as numsp_numlg
numspB	char	How many speaking in interval from overlap tier up to 3+
numlgB	char	How many laughing in interval from overlap tier up to 3+
floorB	char	Combination of numspB and numlgB as numspB_numlgB

Table 5.2 – Variables in Dataset

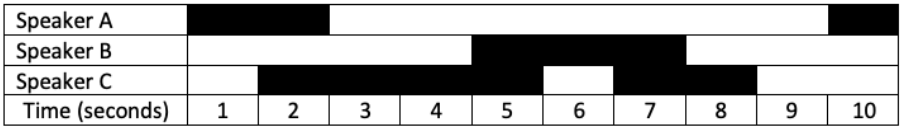


Figure 5.1 – Three participants in a 10-second stretch of conversation, where black represents speech and white represents silence. Speaker A speaks for a total of 3 seconds (2 at the beginning and 1 at the end), Speaker B speaks for a total of 3 seconds, and Speaker C speaks for a total of 6 seconds(4 seconds and 2 seconds). The conversation duration is 10 seconds, while the total participant time is 3+3+6=12 seconds. The total conversation time occupied by speech is 9 seconds. There is global silence for 1 second.

5.3 Participant and Conversation Time

When looking at the composition of conversation in terms of speech, silence, and laughter, there are two points of view possible, giving rise to different statistics. The first way of looking at the conversation is to consider total production by all participants within the conversation, thus ignoring whether any particular speech or laughter is produced 'in the clear' or in overlap. Analyses using this viewpoint are analyses of participant time. The second viewpoint is that of the occupancy of the floor, where at each moment there are a number of possibilities; the conversation is silent, a single participant is speaking or laughing alone, or combinations of two or more participants are speaking or laughing in overlap. Analyses using this viewpoint are analyses of conversation time. Figure 5.1 illustrates the difference between the viewpoints. The use of the Praat speaker tiers and overlap tiers described in Chapter 4 allowed conversations to be considered from both viewpoints.

This chapter has described the methods and measures used in the analysis described in the following chapters.

Part III

Results

Chapter 6

Composition: Corpus, Conversation, and Phase Level

6.1 Introduction

This chapter explores the composition of casual talk at the corpus, conversation, and chat or chunk phase level in order to partially address Research Question 1.

> **What is the compostition of multiparty casual talk, and of its constituent chat and chunk phases in terms of measures of speech, silence, laughter and disfluency?**

The analysis begins with the dataset as a whole - descriptive statistics are given for the distribution of speech, silence, laughter and overlap in the corpus. The distribution of chat and chunk phases and their duration is then analysed, and the focus finally moves to the composition of chunks in terms of speech, silence, laughter, and overlap - first, descriptive statistics are provided, and then these distributions are contrasted for chat and chunk phases.

Knowledge of the general characteristics of the casual conversation data as a whole will provide a baseline for exploration of how these characteristics vary in such con-

versation's constituent chat and chunk phases. Comparison of chat and chunk phase composition will aid in the construction of a quantitative picture of how these phases differ, which will enhance current high level descriptions of the differences. Knowledge, for example, of differences in the proportion of overlap in chat and chunk phases, can be used as a feature in automatically distinguishing between the two, and ultimately generating convincing artificial dialog.

6.2 Descriptive Statistics for CasualTalk Dataset

The final dataset was analysed to describe the nature of the conversations in terms of distribution of speech, laughter and silence. Analysis was carried out in terms of both participant time and conversational time, as described in Chapter 5 Section 5.3.

6.2.1 Speech, Laughter and Silence per Participant per Conversation

At any moment in conversation, a participant can be speaking, silent, or laughing. As described in Chapter 4, Section 4.9, in the current analysis, speech laughs, intervals where a participant's speech was coloured by laughter but still recognisable as speech, were treated as speech and not given a separate label. The dataset contains approximately 6.5 hours (19614.45 seconds) of participant speech. There are approximately 23 minutes (1385.5 seconds) of participant laughter. The distribution of total seconds of participant speech, silence, and laughter per conversation is shown in Table 6.1.

The speech information in Table 6.1 is shown in proportion to conversation length in the stacked barplots in Figure 6.1, where it can be seen that there is no apparent correlation between the amount of speech produced by a participant and the laughter produced - more talkative participants do not appear to laugh more or less than less talkative participants when participation is viewed across the conversation as a whole.

Figure 6.2 shows the contributions of speech and laughter per participant for each

Conv	Part	Gen	Speech	Laughter	Silence	Total
A	a	f	284.36	25.12	3854.51	4164
A	b	m	1049.88	61.46	3052.65	4164
A	c	m	901.93	121.10	3140.94	4164
A	d	m	941.86	141.71	3080.41	4164
A	e	f	472.12	187.61	3504.25	4164
B	f	m	548.82	84.16	4039.35	4672
B	g	f	2091.35	23.65	2557.33	4672
B	h	m	1040.36	33.81	3598.16	4672
C	j	m	543.35	61.74	3772.67	4378
C	k	f	1362.24	32.27	2983.26	4378
C	l	m	771.05	35.48	3571.24	4378
C	m	m	861.63	15.63	3500.50	4378
D	n	m	1117.67	15.44	1871.03	3004
D	o	f	865.36	63.78	2075.01	3004
D	p	f	781.08	32.29	2190.77	3004
E	q	f	423.63	34.19	1611.28	2069
E	r	m	435.81	53.37	1579.92	2069
E	s	m	728.21	9.80	1331.10	2069
E	t	f	302.88	46.76	1719.46	2069
F	u	f	854.71	78.83	3806.97	4740
F	v	m	856.71	92.28	3791.52	4740
F	w	m	839.69	14.09	3886.72	4740
F	x	f	886.82	29.92	3823.76	4740
F	y	f	652.95	91.03	3996.52	4740
Total	s	s	19608.70	1385.51	72345.10	

Table 6.1 – Speech, laughter and silence in seconds per participant per conversation. Conv = conversation, Part = participant, Gen = gender

conversation, measured in seconds, where conversations with the same number of participants are displayed side by side.

Figure 6.3 shows the proportion of speech and laughter in total production per participant. It can be seen that there is no apparent pattern observable, and the proportion of laughter varies.

It should be recalled that the above figures are based on participant time and refer to the total amount of speech and laughter produced during a conversation, some of which is in overlap.

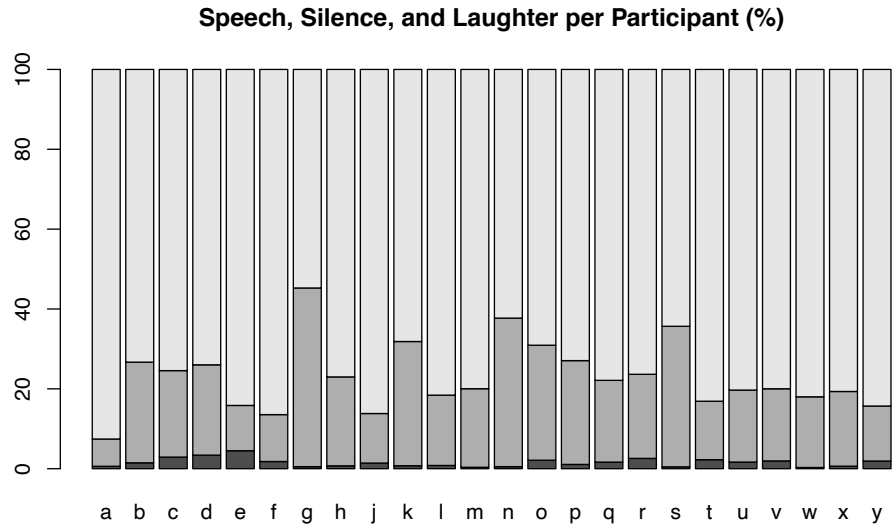


Figure 6.1 – Speech, silence and laughter per participant per conversation. The stacked bars show the amount of speech (dark grey) and laughter (black) produced by each participant, and the time the participant remains silent (light grey), in each conversation, as a percentage of conversation length.

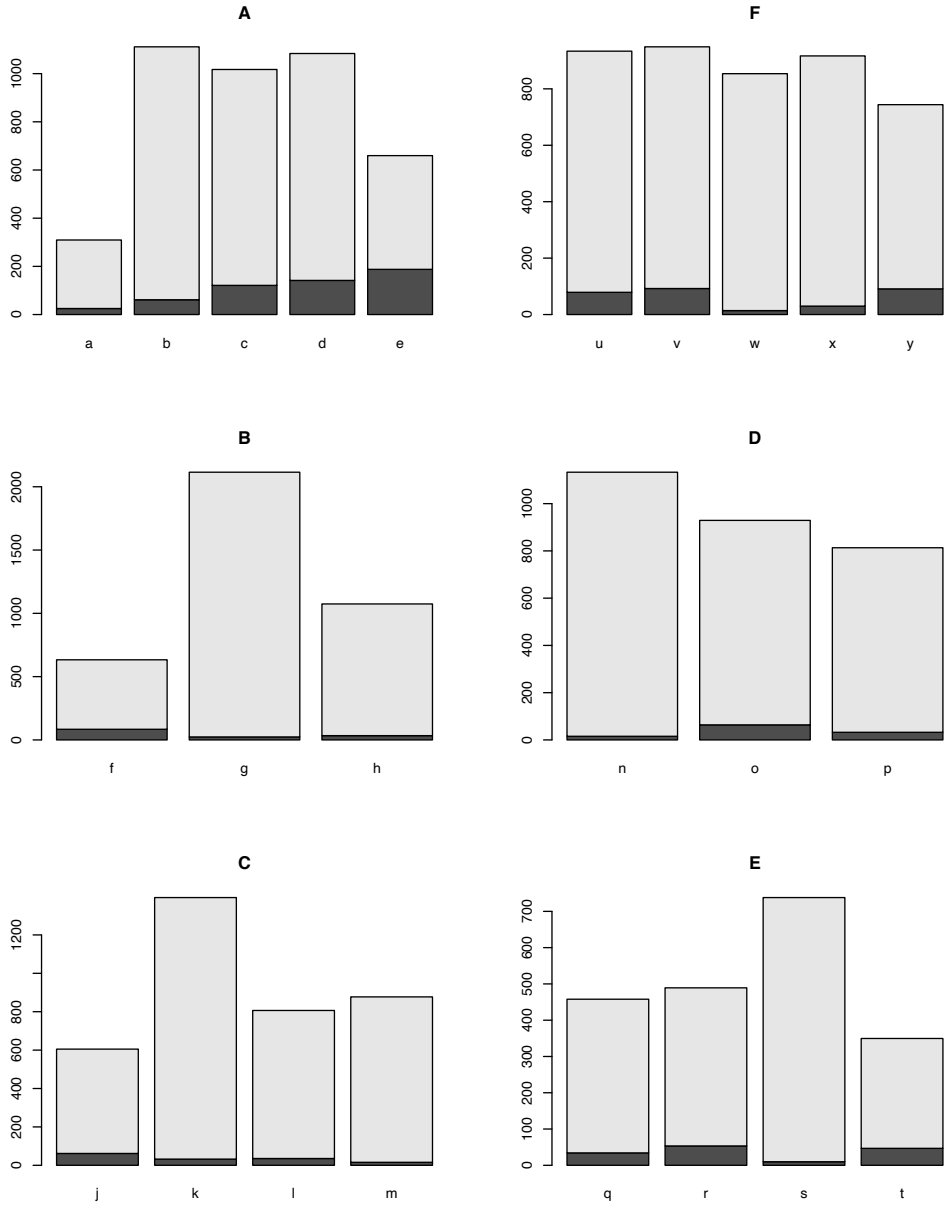


Figure 6.2 – Speech and laughter per participant per conversation. The bar graphs show the amount of speech (grey) and laughter (black) in seconds produced by each participant in each conversation. Conversations with the same number of participants are displayed side by side.

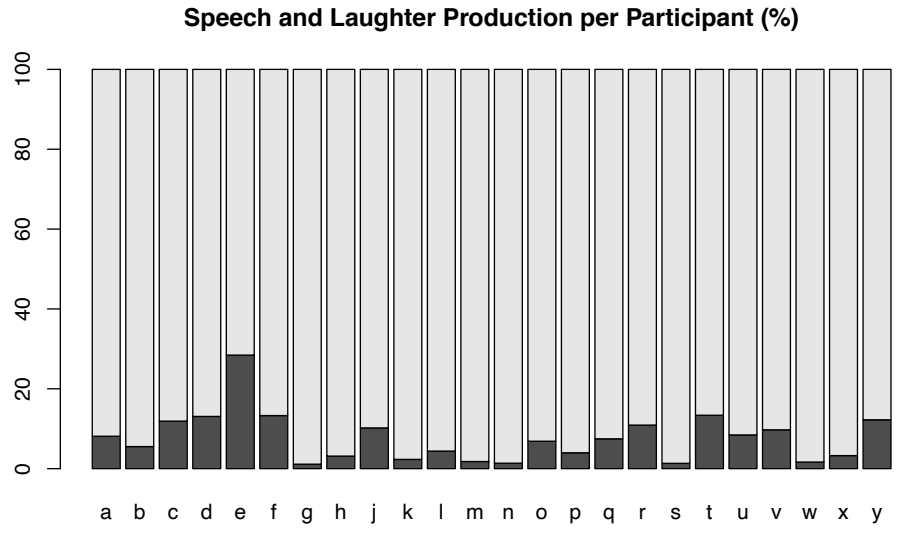


Figure 6.3 – Speech and Laughter production per participant. The bar graphs show the proportions of speech (grey) and laughter (black) as a percentage of total production by each participant

6.2.2 Speech and Laughter Production across Conversations

In order to compare speech and laughter production, the percentage of total participant production (seconds of speech plus seconds of laughter) accounted for by laughter and by speech was calculated. The total speech produced in the dataset was 19614.45 seconds, Total laughter was 1385.513, accounting for 6.6 % of all production. Figure 6.4 shows the proportion of speech and laughter in total production by all participants per conversation. The percentage of production accounted for by laughter ranged from 3.7% to 12.8% with a mean of 6.4%.

These figures should be looked at with caution as conversations from the DANS corpus all contained significantly less laughter than the other conversations in the CasualTalk dataset - this could be because there was a very short approach phase in the DANS conversations compared to those drawn from d64 or from TableTalk, which in turn may reflect the fact that DANS participants were more familiar with one another, or indeed be due to the quality of the recordings, as mentioned in Chapter 4. More generalizable statistics would require a larger corpus.

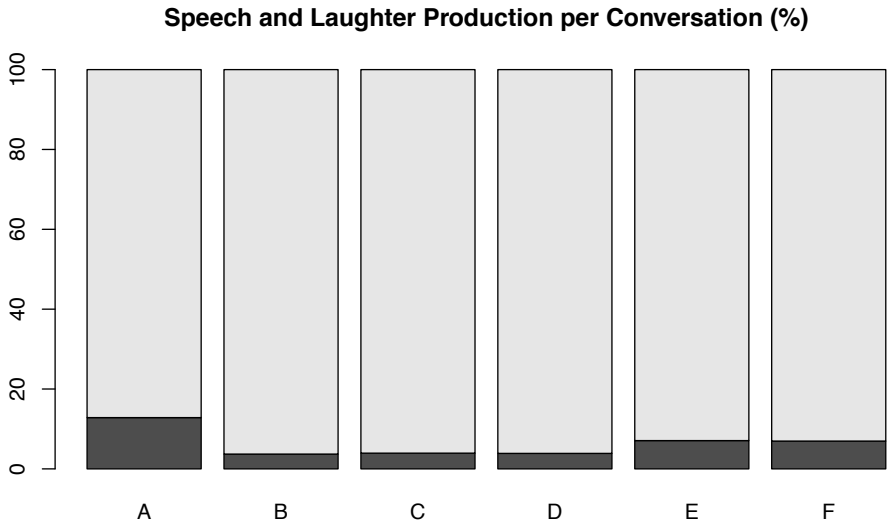


Figure 6.4 – Proportion of speech and laughter in total production per conversation

Looking at these ratios in individual speakers, as shown in Figure 6.3, laughter as a percentage of total production per speaker varied widely, ranging from 1.11% to 28.44% with a mean of 7.66 %.

6.2.3 Conversation Time in Speech, Silence and Laughter

Looking at the conversations in terms of how the conversational floor is occupied over a conversation, there are a number of possible states, depending on how many people are speaking or laughing. In an n-party conversation where laughter is annotated inline with speech, at any instant each participant may be in one and only one of three states - speaking, silent, or laughing. There are thus 3^n possible dialog states at any moment. Table 6.2 shows the time in seconds and the percentage of time spent in various configurations of speech, silence and laughter. For convenience, states where there are 3 or more speakers or 3 or more laughers are grouped as the numbers involved are very small.

Speakers	Laughers	Time(seconds)	Percentage Time
1	0	15219.22	66.09
0	0	928.59	21.40
2	0	1691.00	7.34
1	1	462.12	2.01
0	1	328.06	1.42
3+	0	108.29	0.47
0	2	100.18	0.44
1	2	73.76	0.32
2	1	50.81	0.22
0	3+	42.39	0.18
1	3+	10.78	0.05
2	2	8.25	0.04
3+	1	4.38	0.02
	Total	23027.82	100

Table 6.2 – Conversational time taken up by speech, silence, and laughter in order of prevalence

6.2.4 Overlapping Speech

Disregarding laughter to examine speech overlap alone by treating all laughter labels as silence, there are two possible states for each participant – speaking or silent, and thus 2^n possible states for the conversation to be in at any time, as there are if speech is disregarded in order to examine only laughter. The distributions of conversation time (disregarding laughter) taken up by global silence, single speaker speech, and 2-n speaker overlap for all conversations together and for each individual conversation are shown in Table 6.3 and illustrated in Figure 6.5

.

	Silence	Single Speaker	2-overlap	3-overlap	4-overlap	5-overlap
All	23.42%	68.49%	7.60%	0.47%	0.02%	0.0003%
	5393.47	15771.72	1750.3	107.71	4.56	0.07
A	22.93%	67.52%	8.60%	0.91%	0.05%	0.002%
	954.64	2811.18	357.98	38.20	1.90	0.07
B	24.99%	71.27%	3.71%	0.03%	–	–
	1167.79	3329.79	173.47	1.26	–	–
C	24.28%	70.66%	5.01%	0.045%	0%	–
	1062.95	3093.34	219.50	1.98	0	–
D	15.56%	77.01%	7.27%	0.15%	–	–
	506.38	2313.61	218.49	4.51	–	–
E	24.47%	61.48%	12.38%	1.54%	0.13%	–
	954.64	1272.07	256.12	31.89	2.63	–
F	26.03%	62.26%	11.07%	0.62%	0.0007%	0%
	1234.17	2951.72	524.72	29.86	0.03	0

Table 6.3 – Conversation time taken up in silence, speech, and overlap in percentages and seconds (second line). – denotes impossible condition for conversation - e.g. a 3-party conversation cannot have 4-party overlap

In the corpus as a whole global silence, when all participants are silent, accounts for 23.42% of conversational time, while single party speech accounts for 68.49%. Overlap thus accounts for just over 8% of conversation time. Two speaker overlap is 7.6% while three party accounts for under 0.5%. Overlap involving more than three persons is very rare in the corpus, and can also occur only in 4 of the six conversations,

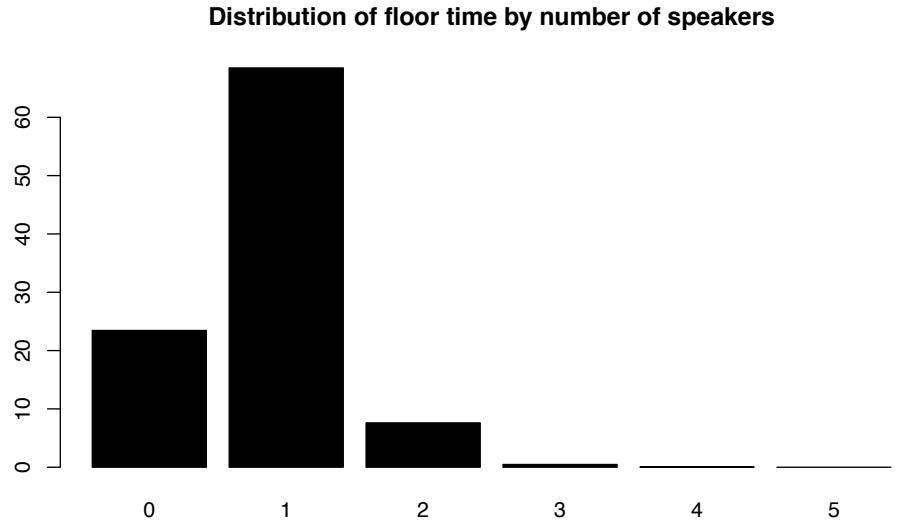


Figure 6.5 – Distribution of conversational floor time in silence (0 speakers) and speech (1-n speakers)

with 0.02% of total conversational time taken up by 4 party overlap (21 instances), while there is only one instance of 5 party overlap in the entire corpus accounting for 70 ms of conversation time. Figure 6.6 shows the proportion of silence, speech and overlap in each of the 6 conversations.

Looking at overlap alone, disregarding one-party speech in the clear and global silence, Figure 6.7 shows the proportional distribution of conversational time spent in overlap by different numbers of speakers. It can be seen that the vast bulk (94%) of overlapping speech involves just one participant overlapping another (2-party overlap).

6.2.5 Distribution of Laughter

There are 1385.51 seconds of laughter in the corpus as a whole, measured on a per participant basis. There are 1080.35 seconds of conversational floor containing laughter - which may be one or more people laughing alone or accompanying another per-

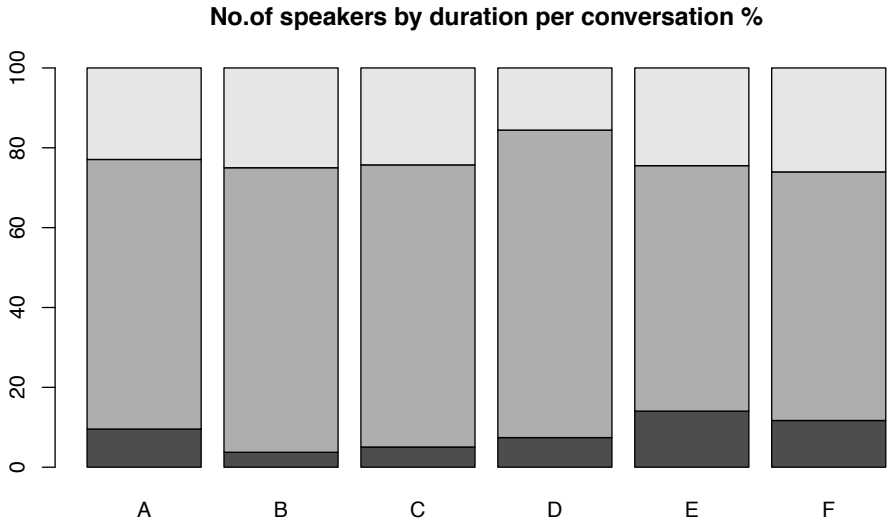


Figure 6.6 – Proportion of overlap (black), speech (grey), and silence (light grey) per Conversation in terms of duration in conversation time

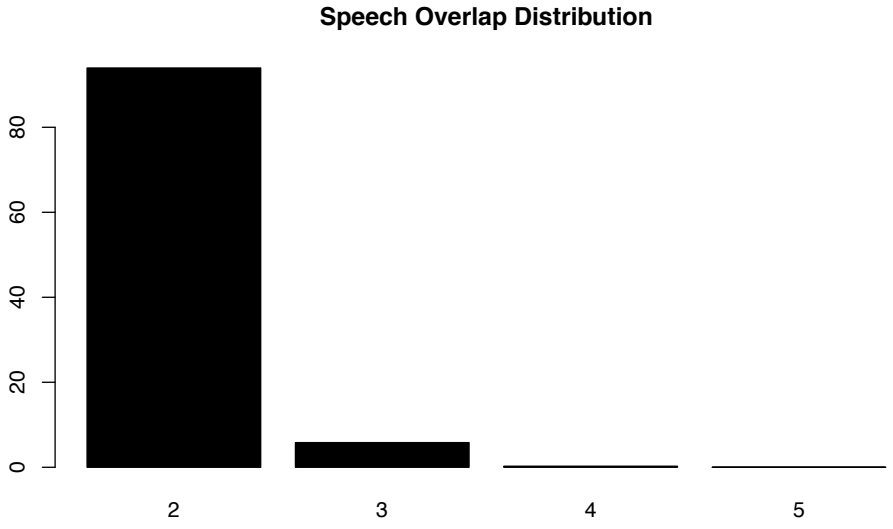


Figure 6.7 – Number of overlapping speakers (% of total duration of overlap)

son's speech. Thus laughter is present in 8.93% of conversational floor time. Table 6.4 shows the distribution of laughter in terms of number of laughers per conversation

by duration, and as a percentage of the total seconds of conversation time containing laughter per conversation.

Conv	1 laugher	2 laughers	3 laughers	4 laughers
A	250.75	86.75	6.58	9.83
A%	67.44	23.33	24.47	2.64
B	116.29	9.64	2.01	
B%	90.89	7.54	0.57	0.00
C	92.94	21.66	1.40	1.16
C%	79.32	18.49	1.19	0.99
D	102.21	4.65	0.00	
D%	95.65	4.35	0.00	
E	99.45	19.05	2.19	0.00
E%	82.40	15.79	1.81	0.00
F	183.73	40.44	6.84	5.25
F%	77.76	17.12	2.90	2.22

Table 6.4 – n-party Laughter in seconds and as percentage of total laughter containing floorspace per conversation

Laughter can occur while another participant or participants are speaking or when there is no speech.

In the corpus, 43.5% (470.6 seconds) of laughter takes place when nobody is speaking, while 56.5% takes place while at least one person is speaking. To further examine how laughter and speech co-occur, Table 6.5 shows the amount and proportion of laughter taking place.

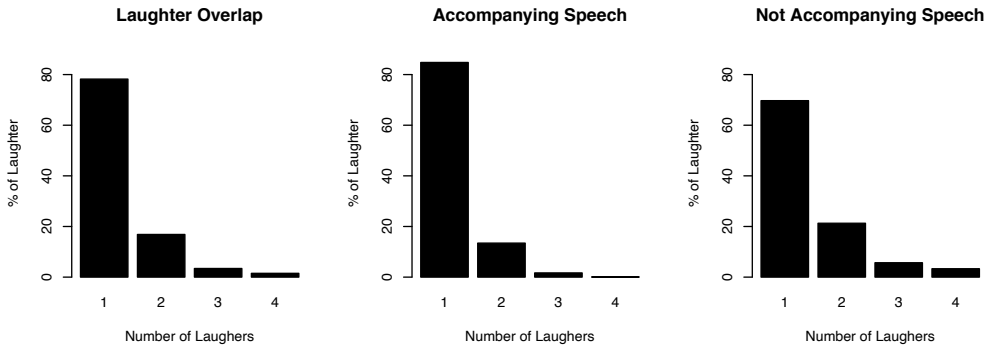


Figure 6.8 – Distribution of laughter overlap with/without speech

Figure 6.8 shows the distribution of laughter by duration across the corpus in terms of how many people are laughing concurrently - overall, in the presence and absence of speech.

Laughers	0 speakers	1 speaker	2 speakers	3 speakers	4 speakers
1	328.06	462.12	50.81	4.00	0.38
	30.36%	42.76%	4.70%	0.37%	0.04%
2	100.18	73.76	8.25		
	9.27%	6.83%	0.76%		
3	26.76	10.16			
	2.48%	0.94%			
4	15.63	0.62			
	1.45%	0.06%			

Table 6.5 – Distribution of laughter occurrence by duration in the presence and absence of talk. Laughter duration, in seconds and percentages of total laughter duration, per number of laughers (participants laughing simultaneously)

It can be seen that the commonest situation for laughter is when one participant laughs while another is speaking (42.76% of all laughter), followed by one participant laughing alone while everyone else is silent (30.36%). Single-party laughter (78.23%) was much more common than shared laughter (21.77%). For shared laughter, 2-party laughter, whether accompanying speech or not, occurs more than three times as often as 3+ party laughter.

6.2.6 Proportion of Speech and Laughter by Gender per Conversation

There were a total of 24 speakers in the conversations - 11 female and 13 male. The total speech by all participants was 19614.45 seconds, of which 46.59% was spoken by female participants. The total laughter by all participants was 1385.513 seconds, of which 45.78% was produced by female participants.

In order to compare speech and laughter production, the percentage of total production (seconds of speech plus seconds of laughter) accounted for by laughter and by speech was calculated. The overall ratio of speech to laughter was almost identi-

cal for each gender with laughter accounting for 6.71% of all production for females, while in males, laughter accounted for 6.50% of all production.

This rather remarkable uniformity in the proportion of speech and laughter in both male and female participants was very close to the 6.6% overall ratio of laughter to total production for all participants. To compare the ratio of single party speech to speech in overlap by gender, the total duration of speech in the clear recorded for each participant was calculated and the percentage of each participant's total speech production which was in the clear (with no accompanying laughter or overlap) was calculated. The mean percentage of clear speech in speech production by females was 73.95% and 75.42% for males. Given the small number of samples, a non-parametric Wilcoxon Rank Sum test was applied to test for significant difference between the distributions of percentage of clear speech in total speech for males vs females. The test showed no significant difference between the distributions (W = 63, p-value = 0.649p).

6.3 Descriptive Statistics for Chat and Chunk Phases

In this section, the characteristics of chat and chunk segments in the dialogues are analysed. As described in Chapter 2 Section 2.3.4, casual conversation has been described as proceeding in phases of interactive talk (chat) or more monologic stretches where a single speaker dominates (chunk). The dataset was annotated for chat and chunk as described in Chapter 4 Section 4.7.

6.3.1 Overall Distribution of Chat and Chunk Segments

There were a total of 571 segments of chat or chunk in the dataset, comprising 213 chat segments and 358 chunk segments. The number and total durations of chat and chunk segments per conversation can be seen in Table 6.6.

It can be seen that in all conversations there were more chunk phases and the time spent overall in chunk phases is greater than that spent in chat phases.

Conv	Chat Number	Chat Duration	Chunk Number	Chunk Duration	Chat/Chunk Ratio(num)	Chat/Chunk Ratio (dur)
A	42	1636.51	73	2527.47	0.58	0.68
B	38	1371.98	82	3300.35	0.46	0.42
C	53	1363.45	68	3014.31	0.78	0.45
D	18	660.68	51	2343.46	0.35	0.28
E	17	909.98	24	1159.13	0.71	0.79
F	45	2168.81	60	2571.70	0.75	0.84

Table 6.6 – Number and duration of chat and chunk segments per conversation

Figure 6.9 shows the proportion of time spent in chat and chunk phases per conversation and overall.

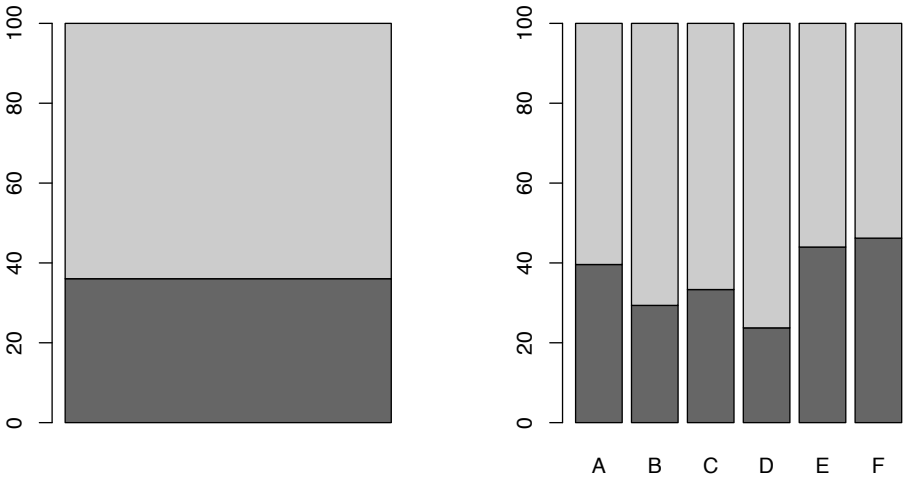


Figure 6.9 – Conversational time spent in chat (dark grey) and chunk phases (light grey) for all conversations (left) and by conversation (right)

6.3.2 Chunk Distribution by Participant

The distribution of chunks per participant per conversation is shown in Table 6.7, and in Figure 6.10. It can be seen that in every conversation, each participant got to take a chunk, but the number of chunks taken by each participant varied widely.

	A	B	C	D	E	F
1	33.33%	58.54%	38.81%	46%	54.17%	30.51%
	24	48	26	23	13	18
2	22.22%	25.61%	26.87%	32%	20.83	20.34
	16	21	18	16	5	12
3	20.83%	15.85%	22.39%	22%	12.5%	18.64
	15	13	15	11	3	11
4	12.5%	–	11.94%	–	12.5	16.95
	9	–	8	–	3	10
5	11.11%	–	–	–	–	13.56
	8	–	–	–	–	8
Total	100%	100%	100%	100%	100%	100%
	72	82	67	50	24	59

Table 6.7 – Chunks per participant by conversation, in percentages and counts (second line). – denotes an impossible condition for a conversation - e.g. a 3-party conversation cannot have a 4th speaker

6.3.3 Chat and Chunk Position

The distribution of chat and chunk phases was analysed to see if the phases were uniformly distributed over the course of a conversation. This was achieved by testing the likelihood of a chunk being followed by a chat or chunk in the first, second, and third ten-minute stretches of each conversation. Thirty minutes was chosen as all conversations exceeded this length by at least ten minutes. Transition or phase change bigrams were collected from the data for chat and chunk. For each chat or chunk phase, a bigram was created linking it to the next phase type – thus bigrams could take the form chat_chunk, chunk_chat, and chunk_chunk. The distribution of phase change bigrams for chunks is shown in Figure 6.11, which graphs the probability of a chunk phase being followed by chat or by chunk throughout the first 30 minutes of conversation. It can be seen that the most likely transition for a chunk in the first ten minutes is to a chat phase, and that the likelihood of chunk-chunk transition grows as the conversation gets longer, with chunk-chunk transitions becoming more and more likely as the conversation continues. Thus, chat and chunk phases are not uniformly

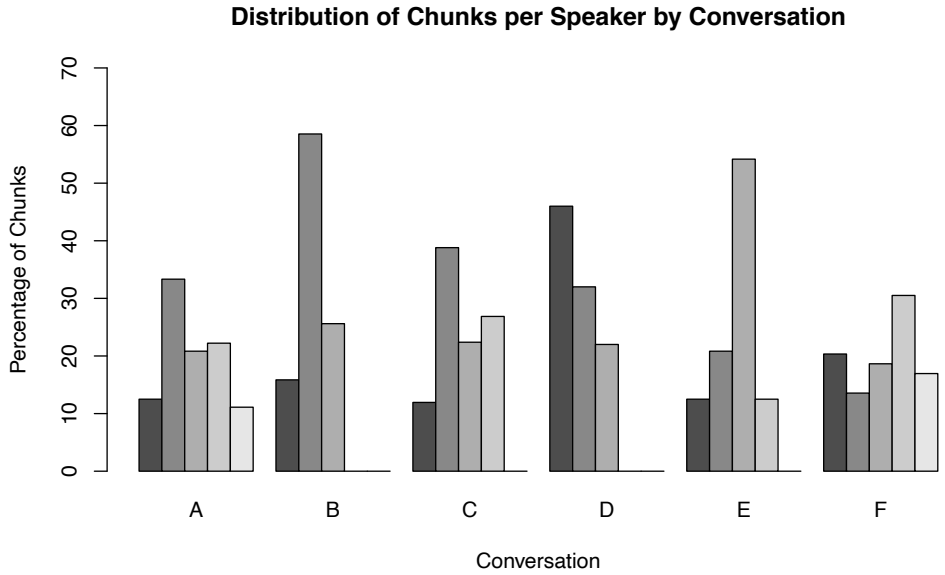


Figure 6.10 – Distribution of chunks per speaker per conversation as percentage of total number of chunks per conversation. Each group of bars represents a conversation (A-F), while individual bars represent speakers.

distributed as conversations progress.

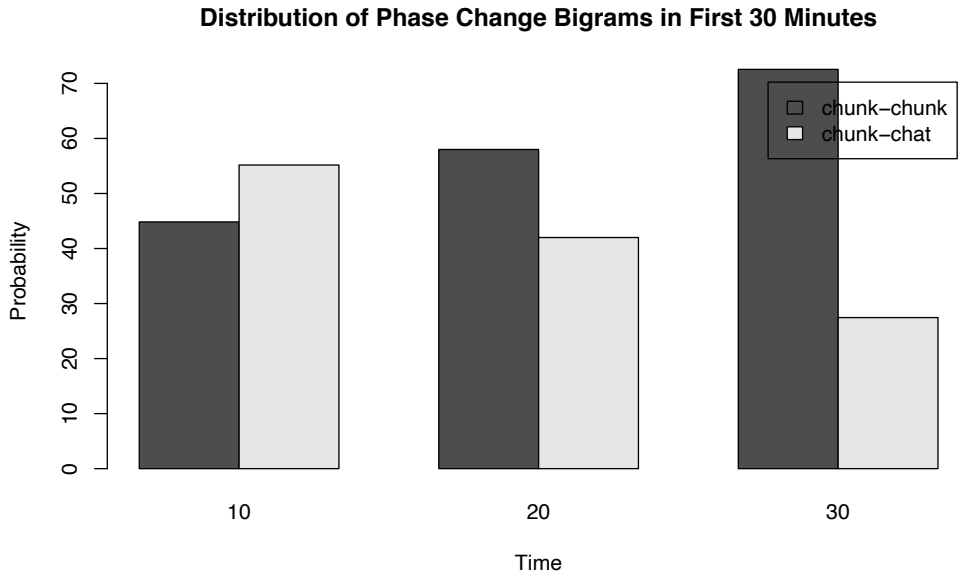


Figure 6.11 – Distribution of chunk-chunk transition (black) and chunk-chat transition (grey) as conversation elapses (x-axis = time) for first 30 minutes of conversation, in 10-minute bins

6.4 Chat and Chunk Duration

The process of comparing distributions for duration is described in some detail here to illustrate how comparisons were made between distributions throughout the thesis.

6.4.1 Mean Durations for Chat and Chunk Phases

Preliminary inspection of the phase (chat and chunk) data showed that the distributions of phase duration in general, and split into chat and chunk, were unimodal but heavily right skewed, with a hard left boundary at 0, as phase durations were by definition always positive. Data of the this type is often closer to a normal distribution when log-transformed and so this transformation is often performed to facilitate modelling. The data were transformed by applying the natural log of the duration, and the resulting distributions are summarised

The chat and chunk phase durations (raw and log) are contrasted in the boxplots in Fig 6.12, where it can be seen that there is considerably more variance in chat durations.

A number of values falling outside the 75th percentile can be seen in the raw data on the left, while there is only one extreme value in the log-transformed data. This value, corresponding to a 287 second long chunk in Conversation C was removed before further analysis was carried out.

The frequency distributions of phase duration for all phases, for chat, and for chunk, are shown in Figure 6.13, where the strong right skew can be clearly seen, while the qq-plots below show that the distributions deviate considerably from normal distributions. In contrast the distributions and qq-plots in Figure 6.14 of log durations for all phases, chat, and chunk can be seen to be considerably closer to normal.

By applying the log transformation skew was reduced from 1.621 to 0.004 for chat and from 1.935 to .0.237 for chunks. These values are below the generally accepted

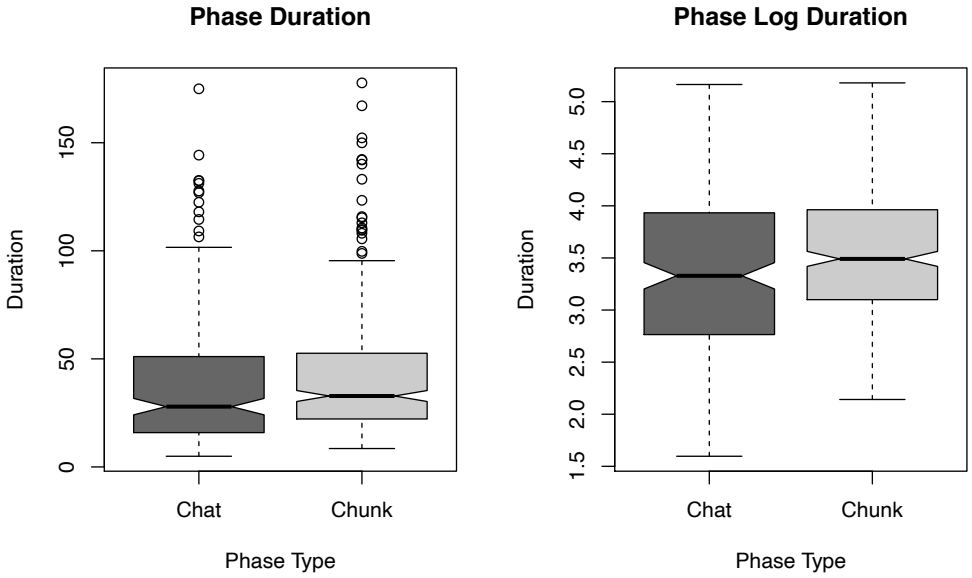


Figure 6.12 – Boxplots of phase duration in chat and chunk phases - raw (left) and log (right)

0.5 threshold for near normality. It was decided to use geometric means to describe central tendencies in the data, after removing one outlying value in the log durations (> 1.5 times IQR).

The geometric means for duration of chat and chunk phases in the dataset were 28.1 seconds for chat and 34 seconds for chunks. These values were close to the median in all cases, in contrast to the elevated mean values seen for the untransformed data.

Wilcoxon Rank Sum tests on raw phase duration data showed significant differences in the distributions of the untransformed durations for chat and chunk (W = 32607, p-value = 0.003191). A Welch Two Sample t-test also showed significant difference in log duration distributions for chat and chunk (t = -3.1309, df = 364.48, p-value = 0.001884).

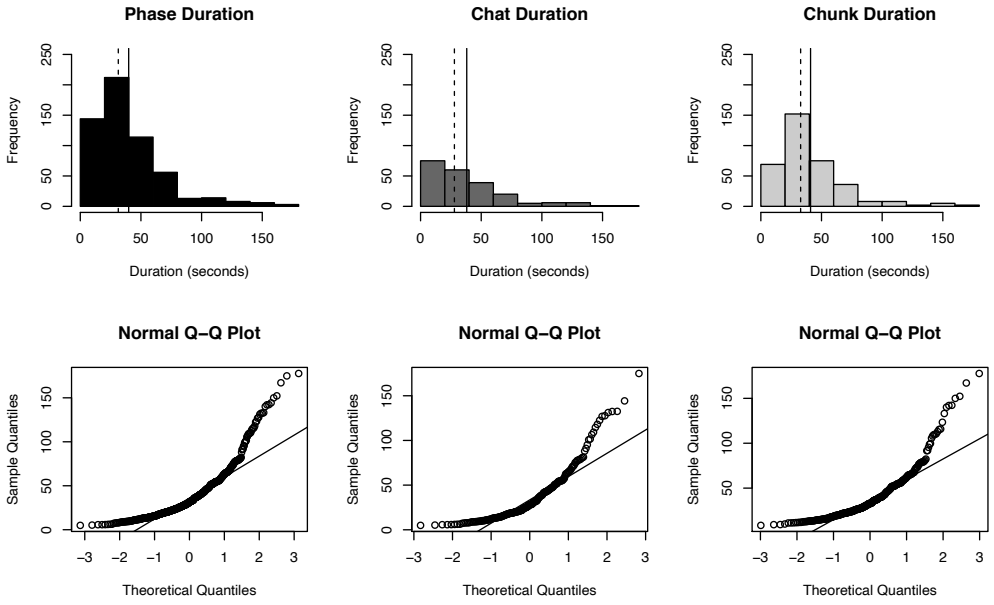


Figure 6.13 – Top: Distribution of the durations of all phases, chat, and chunk phases. Means and medians are marked. Bottom: Corresponding QQ-plots.

6.4.2 Is Chunk Duration Speaker Dependent?

The raw chunk data were checked for speaker dependency using the Kruskal-Wallis rank sum test, a non-parametric alternative to a one-way analysis of variance (ANOVA), and no significant difference in means due to speaker was found (Kruskal-Wallis chi-squared = 34.477, df = 23, p-value = 0.05856), although the p-value was close to 0.05.

6.4.3 Is Chunk Duration Affected by Gender?

Wilcoxon Rank Sum tests on chunk duration data showed no significant difference between duration distributions for chunks owned by male or female participants (W = 17495, p-value = 0.1073)

6.4.4 Is Phase Duration Affected by Conversation?

Kruskal-Wallis rank sum tests on chunk duration showed no significant difference between duration distributions for chunks from different conversations (Kruskal-Wallis

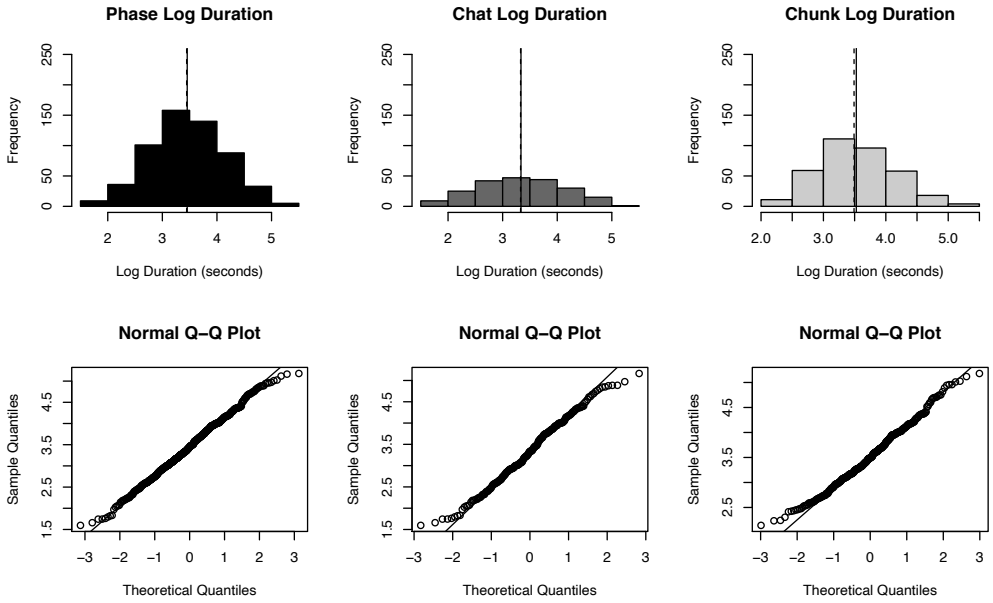


Figure 6.14 – Top: Distribution of the log durations of all phases, chat, and chunk phases. Means and medians are marked. Bottom: Corresponding QQ-plots.

chi-squared = 8.6859, df = 5, p-value = 0.1223). Figure 6.15 shows the distributions of chat and chunk durations and log durations by conversation; the distributions for chunk duration can be seen to be fairly uniform across conversations.

Kruskal-Wallis Rank Sum tests applied to chat duration showed significant differences between duration distributions for chats from different conversations (Kruskal-Wallis chi-squared = 15.617, df = 5, p-value = 0.008027).

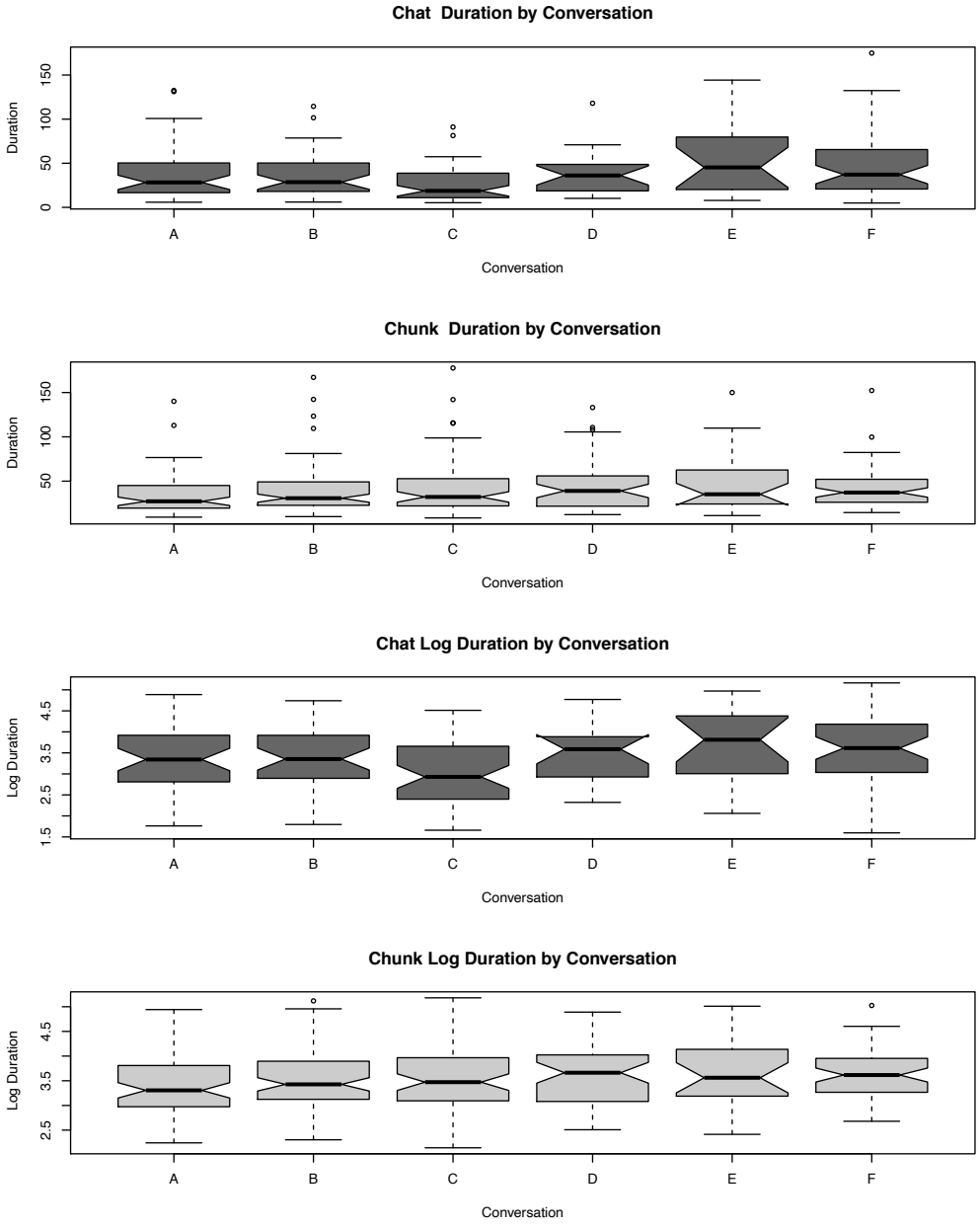


Figure 6.15 – Raw and log durations of chat and chunk phases per conversation.

6.4.5 Does Phase Duration Vary with Distance from Conversation Start?

In order to test whether phase duration correlated to the distance of its start point from the beginning of a conversation, this distance was calculated for each chat and chunk phase. Kendall's Rank Correlation Tau was calculated for both chat and chunk durations and their distance from the start of conversations. The non-parametric Kendall's Tau was used as the duration distributions were non-normal and their log distributions did not approach normality. Kendall's Tau for chat distance from conversation start was -0.00792591 (z = -0.17329, p-value = 0.8624) and 0.0467733 (z = 1.3154, p-value = 0.1884) for chunk distance from conversation start, and no correlation was found for either.

6.5 Chat and Chunk Composition

This section examines how speech, laughter, and overlap are distributed in chat and chunk phases, and also contrasts activity by chunk owners during their chunk phases with other speakers during chunks, and all speakers in chat.

6.5.1 Distribution of Speech and Laughter in Chat and Chunk Phases

Comparing the production by all participants in all conversations, where a participant may produce either laughter or speech, (Table 6.8), laughter accounts for approximately 9.5% of total duration of speech and laughter production in chat phases and 4.9% of total duration of speech and laughter production in chunk phases.

Prod	Chat	Chat(%)	Chunk	Chunk(%)	All	All(%)
Laughter	744.33	9.47	641.18	4.88	1385.51	6.60
Speech	7111.84	90.53	12502.61	95.12	19614.45	96.40

Table 6.8 – Laughter and speech in total production in chat and chunk

Figure 6.16 shows the proportion of speech and laughter in total production by all

participants in chat and chunk segments across the full dataset.

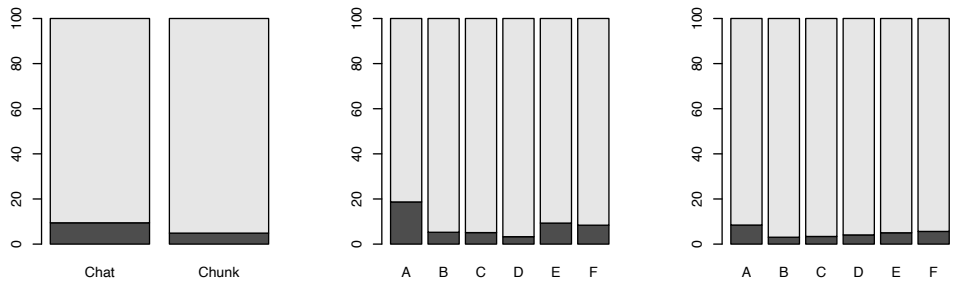


Figure 6.16 – Proportions of speech and laughter production in chat and chunk Phases, across database as a whole (left), in chat by conversation (middle), and in chunks by conversation (right)

In all but one conversation (Conversation D), the overall proportion of laughter vs. speech in chat phases was greater than that of laughter vs. speech in chunk phases, as can be seen in Table 6.9

Conv	Phase	Laughter	Speech
A	Chat	18.75	81.25
	Chunk	8.44	91.56
B	Chat	5.25	94.75
	Chunk	3.05	96.95
C	Chat	5.11	94.89
	Chunk	3.38	96.62
D	Chat	3.04	96.96
	Chunk	4.12	95.88
E	Chat	9.31	90.69
	Chunk	5.02	94.98
F	Chat	8.44	91.56
	Chunk	5.60	94.40

Table 6.9 – Proportion of total laughter and speech production by phase per conversation

Table 6.10 shows the incidence of the different configurations of laughter and speech in chat and chunk phases in the dataset, and the geometric means of the durations of each condition.

Label	Chat %	Chunk %	Chat Dur	Chunk Dur
0_0	28.650	35.498	274	293
0_1	3.862	2.370	191	222
0_2	1.228	0.642	211	186
0_3	0.330	0.166	179	212
0_4	0.101	0.041	470	540
1_0	42.340	45.956	440	664
1_1	3.891	2.825	242	277
1_2	0.869	0.461	195	222
1_3	0.165	0.036	246	155
1_4	0.014	0.005	297	22
2_0	15.772	10.655	267	239
2_1	0.804	0.414	133	161
2_2	0.115	0.026	188	256
3_0	1.687	0.812	163	170
3_1	0.065	0.057	75	130
4_0	0.101	0.031	107	108
4_1	0.007	0.000	383	0
5_0	0.000	0.005	0	70

Table 6.10 – Incidence of silence, speech, and laughter in chat and chunk in CasualTalk. Each line records the incidence (proportion (%) of total floor state labels for the relevant condition and phase) and geometric mean of duration for each condition in chat and chunk. Labels represent how many participants are speaking and laughing - e.g 0_0 represents labels where there is no speech or laughter, 2_1 represents labels where two participants are speaking and one is laughing, etc

It can be seen that global silence happens more frequently in chunks than in chat overall - 35.49% of labels in chunk are global silences compared to 28.65% of labels in chat. A Wilcoxon Rank Sum tests applied to the distributions of silence incidence per chat and per chunk showed that the incidence of global silences was significantly different between chat and chunk phases (W = 23268, p-value = 3.186e-15), with chunk duration significantly higher than chat (W = 23268, p-value = 1.593e-15), with a median percentage global silence of 29.85% in chat and 36.84% in chunk. Wilcoxon Rank Sum Tests showed significant differences in global silence durations in chat and chunk (W = 18600948, p-value = 3.124e-05), with global silences significantly longer in chunk than in chat phases (W = 18600948, p-value = 1.562e-05).

All conditions involving laughter are more frequent in chat than in chunk overall,

Calculating the proportions of laughter production to total (laughter and speech) production per chat or chunk phaseresults in the boxplot shown in Figure 6.17. The distributions are right-skewed with elevated means, so a non-parametric test for significant difference was appropriate. The median for percentage laughter in chunks was 2.2% (mean = 6%) and 5.7% (mean = 9.1%) in chat. Wilcoxon Rank Sum Tests showed significant differences between the proportions of laughter in chat and chunk (W = 45608, p-value = 7.283e-05)., with the proportion of laughter significantly greater in chat (W = 45608, p-value = 3.641e-05).

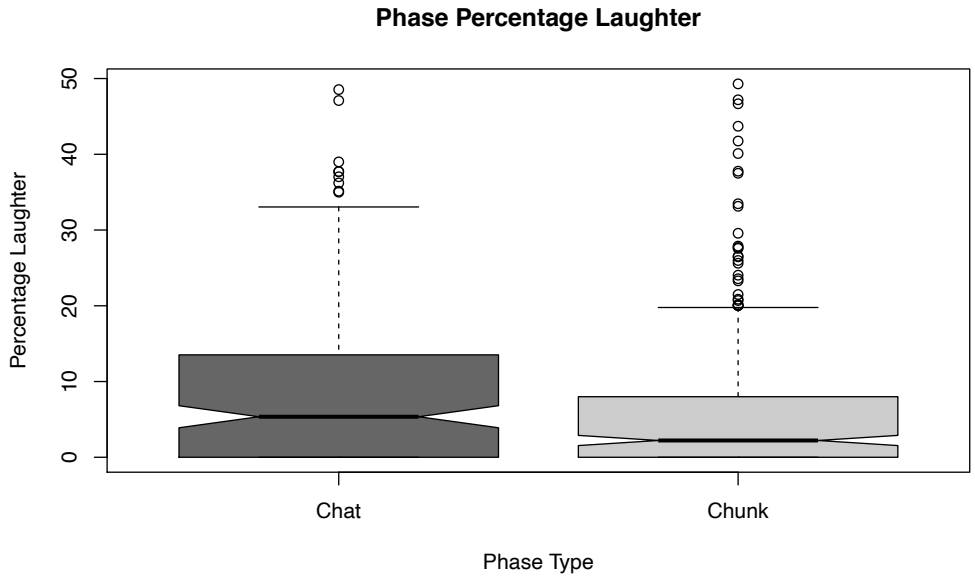


Figure 6.17 – Distribution of proportion of laughter as a percentage of total production of speech and laughter in chat and chunk

6.5.2 Chunk Owner vs Others in Chunks

In the chunks overall, the dominant speakers or chunk owners produced 81.81% (10753.12s) of total speech and laughter, while non-owners produced 18.19% (2390.7s). Table 6.11 shows how much speech and laughter is produced during chunk phases by chunk owners vs all other participants.

	Laughter (s)	Laughter (%)	Speech (s)	Speech (%)
Chunk owners	117.77	18.37	10635.35	85.07
Others	523.41	81.63	1867.26	14.93

Table 6.11 – Total laughter and speech production and in chunks by chunk owners vs other speakers in seconds (s) as proportionally

During chunks, the chunk owners or dominant speakers produced 85.07% of speech, but only 18.37% of laughter, while the trend was reversed for non-owners who produced (as a group) just 14.93% of the total speech but 81.63% of the laughter.

6.5.3 Distribution of Overlap in Chat and Chunk Phases

There is considerable overlapping of speech in the corpora. For the purpose of analysis of speech overlap, laughter was treated as silence and overlap considered as overlapping speech only.

Statistics for overlap were obtained from floor state tiers in the data, with laughter suppressed.

Table 6.12 shows the distribution of conversation time in seconds for all conversations in chat and chunk phases, while Table 6.13 shows the proportion of conversation time spent in each state in percentages. The number of speakers ranges from 0 (global silence), 1 (single speaker), 2 (2 speakers in overlap) to n-1 where n is the number of speakers in the conversation.

From Table 6.12, it can be seen that overlapped speech takes up a far greater proportion of conversation time in chat (17.03%) than in chunk (7.13%), and that overlap

No. Speaking	Chat	Chunk
0	2125.17	3274.05
1	5072.84	10693.03
2	1977.02	773.04
3	61.35	46.68
4	3.29	1.27
5	0.00	0.07

Table 6.12 – Floor occupancy (seconds) in chat and chunk for all conversations

No. Speaking	Chat	Chunk
0	25.76	22.14
1	61.59	72.27
2	11.88	5.26
3	0.73	0.32
4	0.04	0.01
5	0.00	0.00

Table 6.13 – Floor occupancy (%) in chat and chunk for all conversations

involving more than 2 speakers is very rare in both chat and chunk phases. Figure 6.18 shows the proportion of speech, silence, and overlap in chat and chunk in terms of conversation time.

Calculating the proportions of conversation time spent in overlap per chat or chunk phase results in the boxplot shown in Figure 6.19. As with laughter, the distributions are right-skewed with elevated means, although means were less elevated in the overlap than in laughter distributions.

Wilcoxon Rank Sum Tests showed that the proportion of overlapping speech differed significantly between chat and chunk phases (W = 60150, p-value < 2.2e-16), and was significantly greater in chat phases (W = 60150, p-value < 2.2e-16). Wilcoxon Rank Sum Tests showed that the proportion of single-party speech differed significantly between chat and chunk phases (W = 16530, p-value < 2.2e-16), and was significantly higher in chunk (W = 16530, p-value < 2.2e-16).

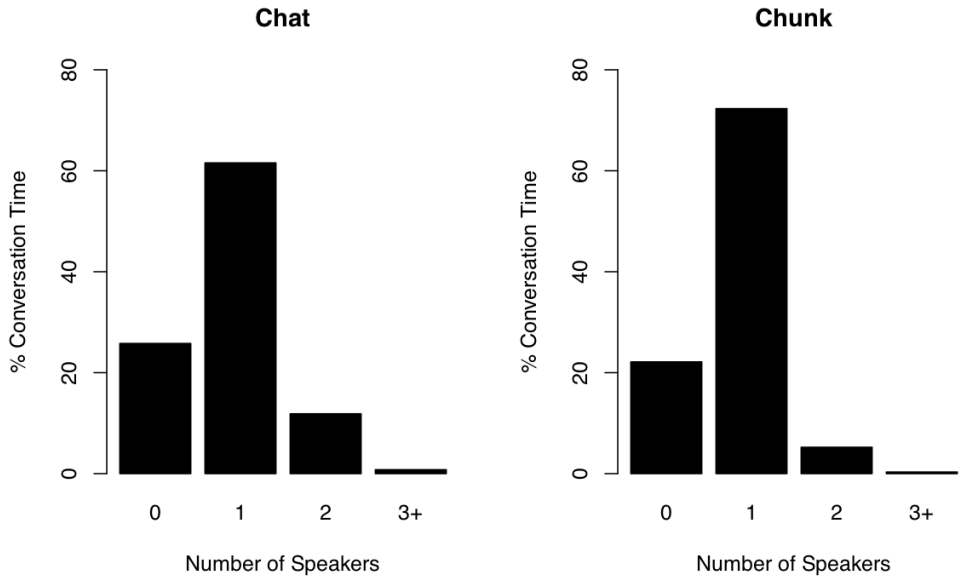


Figure 6.18 – Distribution of silence (0), one-party speech (1), and overlap (2,3+) in chat and chunk phases for all conversations, shown as percentage of total conversation time

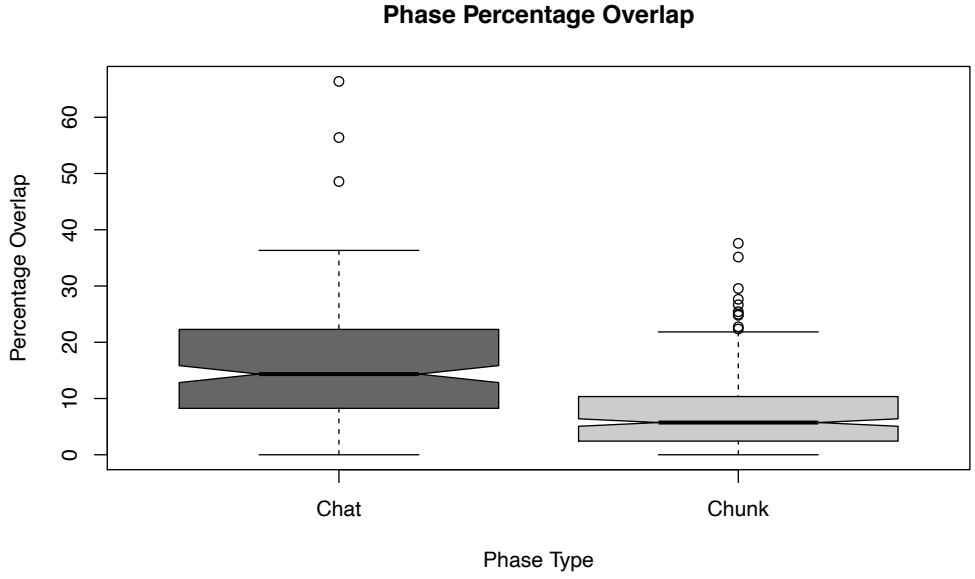


Figure 6.19 – Distribution of Proportion of Floortime Spent in Overlap as a Percentage of Total Floortime spent in Speech in Chat and Chunk Phases

6.6 Conclusion

This chapter has described the dataset in terms of the distribution of speech, silence, laughter, and overlap among participants across conversations as a whole, and in chat and chunk phases. The results of the experiments are further described in Chapter 10. The next chapters describe experiments on conditions around changes in floor state (the number of participants speaking, laughing, or silent at any time) in the conversations and in chat and chunk phases, and the distribution of disfluencies in chat and chunk phases.

Chapter 7

Composition: Disfluency

“To ‘errrr’ is human."

- Shriberg, 2001

7.1 Introduction

This chapter explores the occurrence of disfluencies in Conversation A from the TableTalk dataset. In this chapter, the distribution of disfluencies in Conversation A from the TableTalk dataset is analysed for the conversation as a whole and for chat and chunk phases.

7.2 Disfluency in Conversation A

The audio and text in Conversation A were force aligned using the Penn Aligner and the conversation was then annotated for disfluencies as described in Chapter 4, Section 4.8. The annotated data comprised 42 chat and 73 chunk phases with 14,778 word tokens distributed across 2005 types. Interestingly, UM was the 11th most popular word while UH took 15th place, reflecting the prominence of such tokens in casual talk. The most common word was YEAH.

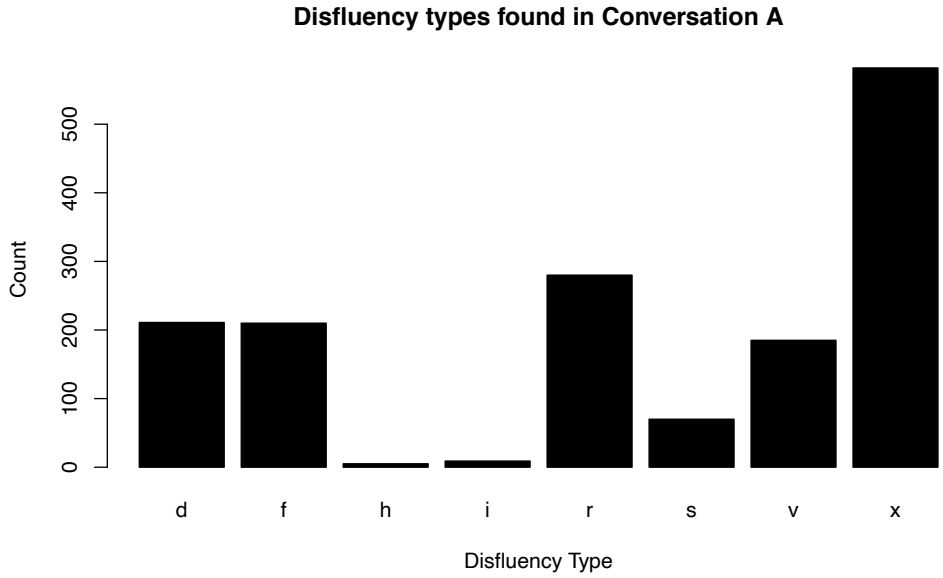


Figure 7.1 – All disfluencies found in data after annotation with Shriberg's Pattern Labelling System (Shriberg, 1994, 53).

7.2.1 Overall Distribution of Disfluency

There were a total of 1552 disfluencies, with total duration of 1016.9 seconds. These disfluencies were divided among 243 distinct labels, of which 16 had ten or more exemplars. These sixteen labels accounted for 1208 or 78% of all disfluencies. The longest label, **[S[R[R[R[r.r].r].r].r]s.xss]** had 28 characters (including brackets), corresponding to the bold text in the phrase:

'it's just **the the the the graphical.. the image** is so beautiful'

The longest time occupied by a disfluent stretch was 4.27 seconds. Table 7.1 shows the number and total duration of spoken utterances and disfluencies per speaker and the proportion of total utterances and disfluencies in the conversation overall contributed by each speaker. The relative disfluency per speaker (DfDur/SpDur) is the percentage of each speaker's total utterance or speech duration occupied by disfluencies, and ranges from 20.1% to 39.2%.

The distribution of all disfluencies including isolated filled pauses and lone pauses

Speaker	a	b	c	d	e
Utterances	401 (7.9%)	1219 (24.1%)	1313 (25.9%)	1353 (26.7%)	773 (15.3%)
Duration (s)	285.11 (7.9%)	1024.93 (28.3%)	900.9 (24.9%)	946.61 (26.1%)	467.67 (12.9%)
Disfluencies	102 (6.6%)	362 (23.3%)	478 (30.8%)	402 (25.9%)	208 (13.4%)
Duration (s)	80.59 (7.9%)	214.92 (21.1%)	353.08 (34.7%)	274.47 (26.9%)	93.88 (9.2%)
DfDur/SpDur	28.2%	20.9 %	39.2%	28.9%	20.1%

Table 7.1 – Total disfluencies per speaker, showing the number and total duration in seconds (s) of utterances and disfluencies per speaker, and their respective percentage proportion in the conversation. DfDur/SpDur shows the proportion of speech occupied by disfluency per speaker.

Disfluency	a	b	c	d	e
Deletions (d)	17 (16.7%)	47 (13%)	67 (14%)	58 (14.4%)	22 (10.6%)
Filled Pauses (f)	17 (16.7%)	42 (11.6%)	91 (19%)	42 (10.5%)	18 (8.7%)
Hybrid (h)	0 (0%)	3 (0.8%)	0 (10%)	1 (0.3%)	1 (0.3%)
Insertions (i)	0 (0%)	5 (1.4%)	3 (0.6%)	1 (0.3%)	0 (0%)
Repetitions (r)	15 (14.7%)	78 (21.6%)	118 (24.7%)	61 (15.2%)	8 (3.9%)
Substitutions (s)	6 (5.9%)	25 (6.9%)	22 (4.6%)	11 (2.7%)	6(2.9%)
Filled Pauses (v)	4 (3.9%)	43 (11.9%)	33 (6.9%)	45 (11.2%)	60 (28.9%)
Pauses (x)	43 (42.2%)	119 (32.9%)	144 (30.1%)	183 (45.5%)	93 (44.7%)

Table 7.2 – Total disfluencies per speaker, showing the breakdown by type of disfluencies per speaker as counts and as percentage of total disfluencies per speaker.

per speaker are shown in Table 7.2 after the Type Classification Algorithm was applied to reduce the number of labels.

7.2.2 Reduced Dataset of Disfluencies

The dataset was edited to further analyse disfluencies. Isolated filled pauses and lone pauses were excluded. As mentioned above isolated filled pauses were marked as **v**; there were 185 in the data, which were excluded, leaving 1367 disfluencies. Lone pauses (x) accounted for 582 labels. These were also excluded, leaving a total of 785 disfluency labels for analysis.

The duration of these 785 disfluencies totals 786 seconds. The total speech by all speakers in the conversation was time was 3625 seconds, so that disfluencies accounted for approximately 22% of the conversation's total speech produced (without adjustment for the time occupied by unfilled pauses within disfluencies).

Speaker	a	b	c	d	e
Utterances	401 (7.9%)	1219 (24.1%)	1313 (25.9%)	1353 (26.7%)	773 (15.3%)
Duration (s)	285.11 (7.9%)	1024.93 (28.3%)	900.9 (24.9%)	946.61 (26.1%)	467.67 (12.9%)
Disfluencies	55 (7.1%)	192 (24.9%)	298 (38.6%)	172 (22.3%)	54 (7%)
Duration (s)	49.49 (7.4%)	143.22 (21.4%)	278.72 (41.6%)	160.49 (23.9%)	38.86 (5.8%)
DfDur/SpDur	17.4%	13.9 %	30.9%	16.9%	8.3%

Table 7.3 – Disfluencies per speaker in the reduced dataset, showing the number and total duration in seconds (s) of utterances and disfluencies per speaker, and their respective percentage proportion in the conversation. DfDur/SpDur shows the proportion of speech occupied by disfluency per speaker.

The TCA grouping of these 785 raw disfluencies (241 labels) resulted in 210 instances of filled pauses, 280 of repetitions, 211 of deletions, 70 of substitutions, 9 insertions, and 5 hybrid cases. For analysis, hybrids and insertions were disregarded.

Table 7.3 shows the number and total duration of spoken utterances and disfluencies per speaker and the proportion of total utterances and disfluencies in the conversation overall contributed by each speaker for the reduced dataset. The relative disfluency per speaker (DfDur/SpDur) is the percentage of each speaker's total utterance or speech duration occupied by disfluencies, and ranges from 8.3% to 30.9%.

The 210 filled pause disfluencies included 113 single filled pauses, with the remaining 94 comprising combinations of filled and unfilled pauses. The 280 repetitions included 137 'pure' repetitions, while the remainder were repetitions with included filled or unfilled pauses.

The 211 deletions included 78 simple deletions. There were 7 deletion/filled pause combinations, 62 deletion and pause combinations, and 19 deletions with filled pauses and pauses. The remaining 45 labels comprised 3 deletions with substitutions, 36 deletions with repetitions, and 6 deletions with editing phrases or discourse markers. The number of deleted words ranged from 1 to 7.

The 70 substitutions included 32 cases of substitution alone and 11 of substitution with filled or unfilled pauses included. The remaining 27 comprised substitutions with repetitions.

Disfluency	a	b	c	d	e
Deletions (d)	17 (30.9%)	47 (24.5%)	67 (22.5%)	58 (33.7%)	22 (40.7%)
Mean Duration (s)	1.15	0.81	0.87	0.98	0.75
Filled Pauses (f)	17 (30.9%)	42 (21.9%)	91 (30.5%)	42 (24.4%)	18 (33.3%)
Mean Duration (s)	0.67	0.62	0.74	0.87	0.43
Repetitions (r)	15 (27.3%)	78 (40.6%)	118 (39.6%)	61 (35.5%)	8 (14.8%)
Mean Duration (s)	0.85	0.70	1.08	0.81	0.86
Substitutions (s)	6 (10.9%)	25 (13%)	22 (7.4%)	11 (6.4%)	6(11.1%)
Mean Duration (s)	0.96	0.97	1.13	1.56	1.29

Table 7.4 – Total disfluencies per speaker in the reduced dataset, showing the breakdown by type of disfluencies per speaker as counts and as percentage of total disfluencies per speaker. The mean duration in seconds (s) of each disfluency type per speaker is also shown.

There were 106 complex disfluencies labelled, 83 with 2 interruption points, 17 with 3, 3 with 4, 2 with 5 and one with six interruption points. Table 7.4 shows the number and total duration of spoken utterances and disfluencies per speaker and the proportion of total utterances and disfluencies in the conversation overall contributed by each speaker for the reduced dataset. The relative disfluency per speaker (DfDur/SpDur) is the percentage of each speaker's total utterance or speech duration occupied by disfluencies, and ranges from 8.3% to 30.9%.

7.3 Disfluency in Chat and Chunk Phases

There were a total of 298 disfluencies during chat phases and 473 during chunk phases. These correspond to three different speaker situations – chunk owner during chunks, non-chunk owner during chunks, and all participants during chat. Table 7.5 shows the numbers of each type of disfluency found in each situation and the relative percentages of each speaker type's total disfluencies.

Disfluencies by speakers other than the chunk owner during chunks total 39, by chunk owners during chunks total 434, while disfluencies by all speakers during chat total 289. In chunks, chunk owners hold the floor with minimal input from other speakers, while in chat all participants can participate equally. In order to contrast

	Chunk Owner in Chunk	Non-chunk Owner in Chunk	Chat
Deletions	101 (23.48%)	14 (%)	96 (32.21%)
Filled Pauses	131 (30.46%)	9 (%)	70 (23.49%)
Repetitions	162 (23.48%)	16 (%)	102 (34.23%)
Substitutions	36 (8.37%)	4 (%)	30 (10.07%)

Table 7.5 – Counts and percentages by speaker type for disfluencies in Conversation A

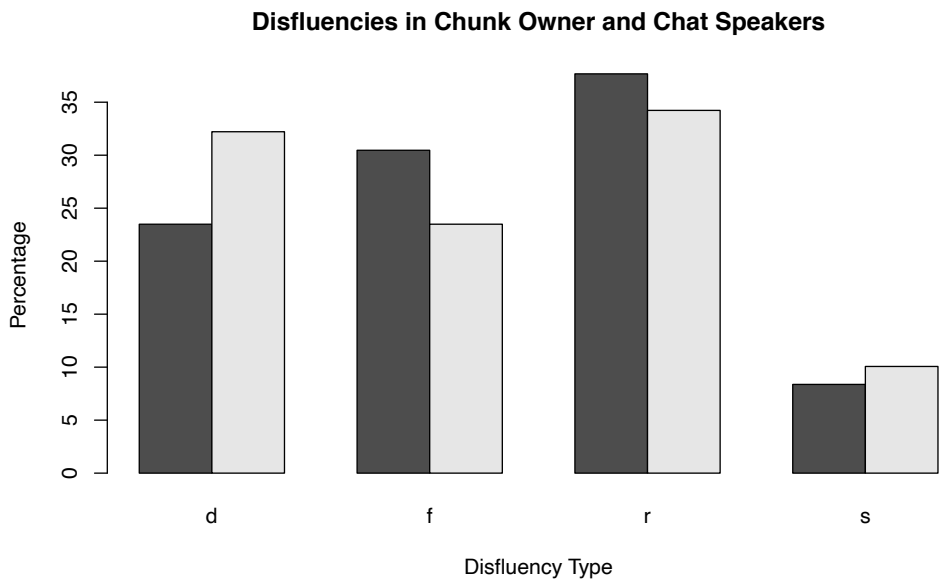


Figure 7.2 – Distributions of disfluencies (deletions, filled pauses, repetitions, and substitutions) in chunk owner speech in chunks (black) and all speakers in chat (grey) as percentage of total disfluencies for each condition

these two modalities, the 39 disfluencies by non-chunk owners during chunks were excluded.

Figure 7.2 shows the distribution of disfluency types (deletion, filled pause, repetition, substitution) occurring in chunk owners speaking during their chunk and by all participants during chat. It can be seen that there is not a great difference in the distributions, but that deletions are proportionally more common in chat phases while filled pauses are proportionally more common in chunk phases overall.

In view of the very small sample of speakers, the distributions for each speaker

were examined to see if the differences observed overall were present for all speakers. Although the proportions varied for individual speakers, in all cases, filled pauses were also proportionately higher in chunk owner speech versus other speech, and repetitions were also proportionally higher in chunk owner speech for each speaker.

The mean durations of each disfluency type per speaker were extracted for the three conditions and no pattern of differences was observed - some speakers had longer deletion disfluencies in chat than as chunk owners while the situation was reversed for others.

7.3.1 Disfluency Rates

Two measures of disfluency in speech were calculated. First, the proportions of disfluent speech to total speech by duration in each speaker overall, in chat, as chunk owner, and as non-owner speaking in chunks, were calculated to give a measure of the proportion of disfluent speech by each speaker. Second, the number of disfluent stretches per second of spoken production was calculated per speaker in all of the above conditions. Table 7.6 shows the results of these calculations. There was no pattern observed of disfluency rate differences with different roles among the different speakers.

7.3.2 Filled Pauses in Chat, Chunk Owner and Non-Chunk Owner Speech

Filled pauses in chat, chunk owner and non-chunk owner speech were gathered to investigate if there were any differences in their duration based on speaker role. As described in Chapter 4 Section 4.8.3, filled pauses were marked as **f** when they occurred within or adjacent to an utterance by a speaker, i.e. as part of a speaker's turn or attempt to take a turn. Isolated 'uh' and 'um' filled pause like productions were marked **v**, but were not considered disfluencies for the analyses above.

The dataset contains 185 disfluencies marked as isolated filled pauses (**v**). There

Speaker	a	b	c	d	e
C df dur	8.73	60.83	93.67	57.54	30.52
CO df dur	39.81	71.02	164.61	94.82	
4.5					
NCO df dur	0.95	11.37	20.43	8.14	3.84
C sp dur	82.4	396.76	312.89	387.97	261.63
CO sp dur	192.83	566.06	541.84	464.03	131.78
NCO sp dur	9.87	62.11	46.17	94.61	74.26
C df	12	77	98	72	39
CO df	42	102	188	89	9
NCO df	1	13	12	11	6
C %df	10.59	15.33	29.93	14.83	11.66
CO % df	20.64	12.55	30.38	20.43	3.42
NCO % df	9.58	18.31	44.26	8.6	5.17
C dfps	0.146	0.194	0.313	0.186	0.149
CO dfps	0.217	0.18	0.347	0.192	0.068
NCO dfps	0.101	0.209	0.259	0.116	0.081

Table 7.6 – Duration of disfluency and speech and number of disfluencies per speaker (a to e) when speaking as Chunk Owner in Chunk (CO), Non-Chunk Owner in Chunk NCO), and in Chat(C). Proportion of Disfluency (%df) and Rate of Disfluency (dfps) in Chunk Owners in Chunk (CO), Non-Chunk Owners in Chunk NCO), and in Chat(C)

	Chunk Owner in Chunk	Non-chunk Owner in Chunk	Chat
f-type	63	5	45
mean duration (ms)	337	259	257
v-type	0	117	68
mean duration (ms)	–	292	305

Table 7.7 – Distribution of filled pauses by type and by role

were 210 **f** type filled pauses, of which 87 were filled pauses combined with silences. The combined pauses were removed leaving 113 f-type filled pauses.

The remaining **f** and **v** filled pauses were gathered together, and grouped by type and speaker role as shown in Table 7.7 where counts and mean durations are shown for the various combinations. The mean duration of filled pauses in chunk owner speech is longer than any of the other conditions. The mean durations of filled pauses in the vicinity of utterances by the same speaker is very similar for non-chunk owners in chunk and for speakers in chat. These f-type pauses for everyone but chunk owners are shorter than isolated pauses produced in chat or by non-chunk owners in chunks.

However, these results are based on small numbers of speakers and whether isolated pauses are shorter in each speaker varies.

7.4 Conclusions

This chapter has reported experiments analysing the occurrence of disfluencies in Conversation A of the CasualTalk dataset. The results are discussed in Chapter 10

Chapter 8

Dynamics: Changes in Floor State

"Conversation should be like juggling; up go the balls and plates, up and over, in and out, good solid objects that glitter in the footlights and fall with a bang if you miss them."

- Waugh, 1945

8.1 Introduction

In this chapter, analysis focusses on conversation dynamics at the local utterance or turntaking level, in order to address Research Question 2.

What are the characteristics of casual talk, and its constituent chat and chunk phases, in terms of floor state dynamics at the utterance level?

The nature of casual talk is examined in terms of the durations of individual IPUs or utterances, in speech and laughter, and the duration of stretches of overlapping speech. The level of 'liveliness' or interaction is measured using two methods and the liveliness of chat and chunk talk compared. The mechanics of casual talk are explored using representations of the floor state, which provides an overview of who is speaking at any time in the conversation, and changes with the onset and offset of

solo speech, overlap, or silence. Throughout a conversation, floor state changes occur constantly, as speakers start and stop speaking, solo or in overlap, and general silences occur. These changes are explored for the entire dataset and for chat and chunk phases. The mechanics of casual talk are explored in terms of within and between speaker floor state transitions, analogous to turn retention and change.

The duration of speech, silence and overlap intervals, and the type of overlap encountered, in chat and chunk phases are contrasted. Then, the level of interactivity in casual talk is explored using two measures, an existing compression measure and a novel floor change rate measure. Liveliness levels are contrasted in chat and chunk phases. The environment around silences, overlap and speaker changes is then investigated more fully, in order to better understand what happens around speaker changes where there are more than two participants in a conversation. Within and between speaker transitions, the evolution of the conversational floor state from a stretch of single party speech in the clear to the next stretch of single party speech in the clear by the original or a different speaker, are investigated and analysed across whole conversations. The analysis of within and between speaker transitions then continues at the level of individual chat and chunk phases.

8.2 Speaker Contributions, Utterance and Laugh Duration in Chat and Chunk

The distributions of interpausal unit (IPU) lengths for speech and laughter in terms of participant time spent – the length of individual utterances by speakers in chat and chunk – were contrasted by categorising utterances according to situation and role of participant. Three classes were created - Chunk Owner Speaking in Chunk covering all activity by the owner of a chunk during that particular chunk, Non-Chunk Owner speaking in Chunk covering all speech by anybody other than the chunk owner during a chunk, and Chat Speaker covering all activity during chat phases.

	Chunk Owner in Chunk	Non-chunk Owner in Chunk	Chat Speaker
Speech	1691	611	1138
Laughter	761	907	918

Table 8.1 – Mean duration in milliseconds of IPUs and laughs in Chunk Owner, Non-Chunk Owner, and Chat Speakers

IPU length for speech and laughter is shown in Table 8.1 where utterance means are shown for Chunk Owners (speaking in chunks), Chat Speakers (all speakers in chat), and Non-Chunk Owners (speakers other than chunk owner speaking in chunks). The mean duration of utterances by chunk owners is greater than that of chat speakers, while both are greater than the duration of utterances by non-chunk owners in chunks. Wilcoxon Rank Sum Tests showed significant differences in utterance length between Chunk Owners and Chat Speakers (W = 25741542, p-value < 2.2e-16), Chunk Owners and Non-Chunk Owners (14488990, p-value < 2.2e-16), and Chat Speakers and Non-Chunk Owners (W = 12171445, p-value < 2.2e-16). Chunk Owner utterances were significantly longer than both Chat Speaker utterances (W = 25741542, p-value < 2.2e-16) and Non-Chunk Owner utterances (14488990, p-value < 2.2e-16), while Chat Speaker utterances were also significantly longer than Non-Chunk Owner utterances (p<1.2e-16).

For laughter, Wilcoxon Rank Sum tests showed significant differences in laughter duration between Chunk Owners and Chat Speakers (W = 53715, p-value = 0.00612), and Chunk Owners and Non Chunk Owners (W = 37921, p-value = 0.0103). Chunk Owner laugher vocalisations were significantly shorter than Chat Speaker laughs (W = 53715, p-value = 0.00306), and also significantly shorter than Non-Chunk owner laughs (W = 37921, p-value = 0.005148). There was no significant difference in laugh length between Chat Speakers and Non-Chunk owners.

The duration of two-party intervals of overlap were compared in chat and chunk phases, as two-party overlap is by far the most common overlap type in the data.

Wilcoxon Rank Sum tests showed significant differences in two-party overlap duration in chat and chunk phases (W = 2332703, p-value = 0.0002105), with two-party overlap durations significantly higher than in chat phases (W = 2332700, p-value = 0.0001053). Wilcoxon Rank Sum Tests on global silence length in chat and chunk phases showed that silences in chunk phases were significantly longer than in chat phases (W = 18601000, p-value = 1.562e-05), as seen in Chapter 6 Section 6.5.3)..

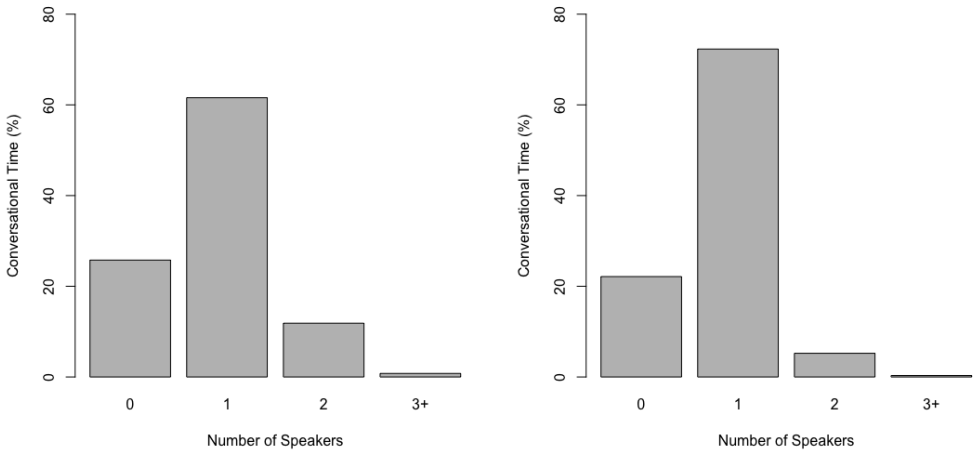


Figure 8.1 – Distribution of the floor in terms of % duration in chat (left) and in chunk (right) phases. X-axis shows number of speakers (0,1,2,3+) speaking concurrently

8.3 Floor State Change and Speaker Activity

As discussed in Chapter 2, Section 2.4, the use of floor state allows conversations to be analysed in terms of intervals where a particular speech/silence configuration or floor state holds, rather than more theory-dependent utterances or turns. As described in Chapter 4, Section 4.5, the Praat annotations were used to generate a floor state tier in Praat, which provided labels for the conversational floor state throughout the conversations.

An extract from the original Praat textgrid from a chat segment in Conversation

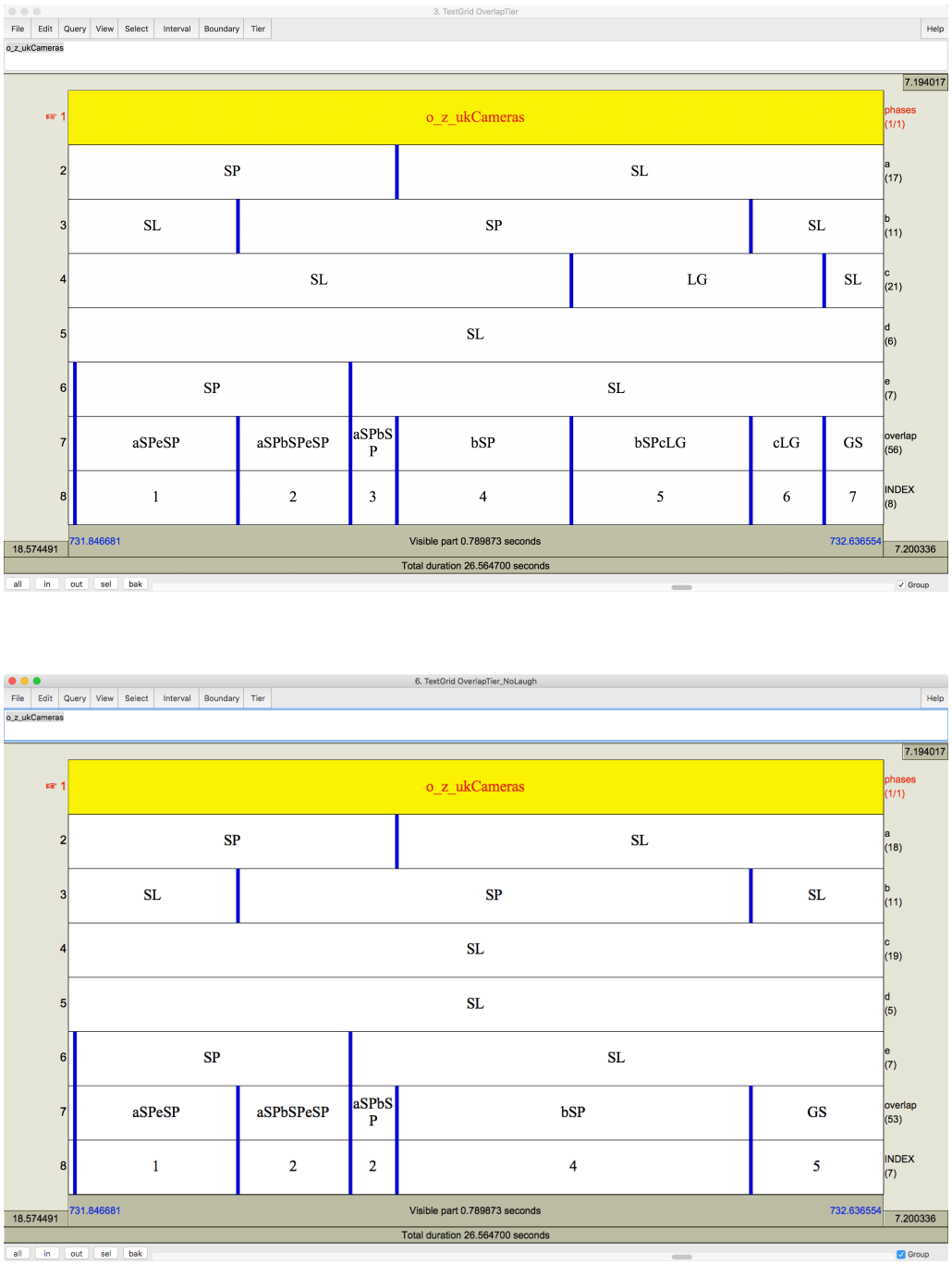


Figure 8.2 – Extract from Praat textgrids for Conversation A, showing the same segment of conversation, part of a chat phase, annotated for phase (Tier 1), speech (SP), laughter (LG), and silence (SL) from participants a-e on five participant tiers (Tiers 2 to 6), with the generated floor state tier (Tier 7), and an index tier (Tier 8) indexing the floor state labels with a single digit for convenience. The top textgrid shows the segment with laughter included (top), laughter is excluded in the bottom segment.

A is shown in the top part of Figure 8.2. In the extract, which is just over 0.8 seconds long, five participants (a-e) are involved in a chat segment on the subject of cameras in the UK. In Tier 7, the floor state (overlap) tier, on the top text grid there are seven different floor states in succession, each with different combinations of speech, silence, and laughter from the participants. As an example, in segment 2, **aSPbSPeSP**, speakers **a**, **b** and **e** are talking, while in segment 4, **bSPcLG**, speaker **b** is talking and **c** is laughing, and in segment 7, **GS**, there is global silence. In an n-party conversation with a label set consisting of speech, silence and laughter, there are 3^n possible speaker states; 243 in this 5-party case.

For the purposes of the analysis below, laughter was disregarded by treating it as silence and the Praat tiers were adjusted, reducing possible speaker states to 2^n, or 32 in this 5-party example. The same segment from Conversation A taken from the adjusted Praat tier, without laughter, is shown in the bottom part of Figure 8.2, where it can be seen that the number of floor states has reduced to five successive floor states from the original seven.

8.4 Conversational Liveliness

The level of activity or liveliness of conversations in the CasualTalk dataset and their constituent chat and chunk phases was analysed using two measures – a novel floor change rate measure and an existing compression measure (Burger et al., 2008) described in Chapter 2, Section 2.4.3. Floor change rate is measured as the number of changes in floor state per second in a stretch of interaction, and floor change rate per second per person is the floor change rate divided by the number of participants in the interaction. Compression is defined for speech as the total duration of speech in a stretch of interaction (the sum of speech durations in participant time) divided by the conversation time occupied by speech, and thus takes a value greater than equal to one for any stretch of interaction, with higher values indicating greater incidence of

simultaneous speech. Compression can also be calculated for laughter and for speech and laughter combined.

The measures were contrasted in chat and chunk phases, exploring the level of liveliness in the different phases and testing the hypothesis that chat phases are more lively than chunk phases.

8.4.1 Floor State Changes and Floor State Change Rates

The CasualTalk dataset contains 30688 changes in floor state when speech and silence only are considered, over a total of 23030 seconds, an average of 1.3 per second.[1]

Table 8.2 shows the total durations of chat and chunk phases overall and by conversation, and the number of floor state changes (intervals on the floor state tier) by chat and by chunk. From these, two floor state change rates are derived - the floor state change rate per second (FCs) and the floor change rate per second per speaker (FCsp). The floor state change rate per second (FCs) is defined as by the number of floor state changes occurring during the period of interest divided by the duration of the period of interest . The second measure, floor state change per second per person (FCsp) is defined as FCs divided by the number of participants in the interaction.

It can be seen that both per second and per second per person floor change rates are higher for chat than for chunk phases overall. To investigate whether these differences were significant, floor change rate per second (FCs) and floor change rate per second per person (FCsp) were then measured on a chunk by chunk and chat by chat basis, and the distributions of the individual scores in chat and chunk phases were compared.

FCs in chat ranged from 0.6719 to 3.011 changes per second with mean of 1.5461 (median - 1.5047, IQR - 1.2425 to 1.8193), while in chunk the range was from 0.2935 to 2.3462, with mean of 1.2560 (median - 1.2262, IQR - 0.9876 to 1.4885). Wilcoxon Rank

[1] With laughter included, the number of floor state changes is 33251, giving an average of 1.4 per second. The small difference reflects the relative scarcity of laughter in the corpora when compared with speech.

	Overall	A	B	C	D	E	F
Dur(O)	8236	1668	1372	1338	718	918	2195
Dur(X)	14790	2494	3300	3011	2286	1151	2546
FC(O)	12607	2820	1732	1677	963	1707	3708
FC(X)	18081	3445	3401	3311	2387	1938	3599
FCs (O)	1.53	1.69	1.26	1.25	1.34	1.86	1.69
FCs (X)	1.22	1.38	1.03	1.13	1.04	1.68	1.41
No. Sp	–	5	3	4	3	4	5
FCsp(O)	–	0.338	0.42	0.313	0.447	0.465	0.338
FCsp(X)	–	0.276	0.343	0.282	0.347	0.42	0.282

Table 8.2 – Floor state changes overall and per conversation, in chat (O) and in chunk (X) phases, per second (FCs), and per second per person (FCsp)

Sum tests showed significant differences in floor change rate for chat and chunk (W = 53772, p-value = 3.74e-16), and the rate for chat was found to be significantly greater than for chunk (W = 53772, p-value < 2.2e-16).

Looking at FCsp for chat, the range was from 0.168 to 0.7192 with mean 0.3082 (median - 0.3683, IQR - 0.3165 to 0.4303), while the FCsp rate in chunk ranged from 0.05871 to 0.66165 with mean 0.32144 (median - 0.31943, IQR - 0.26023 to 0.37560). Wilcoxon Rank Sum tests showed significant differences in floor change rate per person for chat and chunk (W = 51475, p-value = 3.853e-12), with chat having significantly higher floor change rates (W = 51475, p-value = 1.927e-12).

8.4.2 Compression

As described in in Chapter 2, Section 2.4.3, compression is a measure of the amount of overlap in a conversation, and is defined as the total contributions by all participants in the conversation divided by the amount of time during the conversation devoted to these contributions – that is the total individual contributions without considering whether they are in the clear or overlapped (so each interval of overlapped speech is counted as many times as there are overlapping speakers) divided by the conversation time used for these contributions (so overlapped speech is counted only once) of the

conversation. Figure 8.3 illustrates this concept.

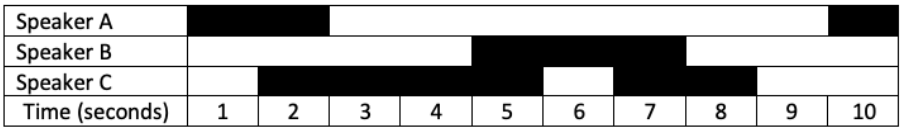


Figure 8.3 – Three participants in a 10-second stretch of conversation, where black represents speech and white represents silence. Speaker A speaks for a total of 3 seconds (2 at the beginning and 1 at the end), Speaker B speaks for a total of 3 seconds, and Speaker C speaks for a total of 6 seconds(4 seconds and 2 seconds). The conversation duration is 10 seconds, while the total participant time is 3+3+6=12 seconds. The total conversation time occupied by speech is 9 seconds. There is global silence for 1 second. Compression for this stretch is therefore 12/9 = 1.33.

Compression is used in the following experiments to measure the level of overlap of speech and of laughter in the corpus as a whole, in individual conversations, and for individual participants. It is also used to compare the level of overlap in chat vs chunk phases, and to compare the levels found for the casual conversation data examined here with those of seminar and meeting data reported by other researchers.

Compression across Conversations

As a first measure, the sum of per participant totals of speech and laughter are divided by the total floor time. This gives an overall speech and laughter compression ratio (ComSL) of 1.16.

The overall compression rate for speech was then calculated as the sum of per participant totals of speech divided by the total floor time. This gives an overall speech compression ratio (ComS) of 1.1. The same procedure for laughter gave a laughter compression rate (ComL) overall of 1.28.

Table 8.3 shows duration and compression values for speech and laughter, speech only, and laughter only for the CasualTalk dataset as a whole and for each of the included conversations. Table 8.4 shows duration and compression values for speech

and laughter, speech only, and laughter only for chat and chunk phases in the CasualTalk dataset as a whole and for each of the included conversations.

	Overall	A	B	C	D	E	F
DurPSL	20994	4181	3822	3683	2876	2035	4397
DurFSL	18099	3351	3561	3380	2572	1618	3617
ComSL	**1.16**	**1.25**	**1.07**	**1.09**	**1.12**	**1.26**	**1.22**
DurPS	19609	3664	3681	3538	2764	1891	4091
DurFS	17629	3204	3505	3315	2537	1563	3506
ComS	**1.11**	**1.14**	**1.05**	**1.07**	**1.09**	**1.21**	**1.17**
DurPL	1385	537	141	145	112	144	306
DurFL	1081	372	128	117	107	121	236
ComL	**1.28**	**1.44**	**1.1**	**1.24**	**1.05**	**1.2**	**1.3**
No. Sp	–	5	3	4	3	4	5

Table 8.3 – Compression of speech and laughter (ComSL), speech only (ComS), and laughter only (ComL) in dataset. Compression is calculated overall and per conversation, using durations of participant contributions for speech and laughter (DurPSL), speech only (DurPS) and laughter only (DurPL) and floor durations for speech and laughter (DurFSL), speech only (DurFS), and laughter only (DurFL). The number of speakers (No. Sp) in each conversation is given for reference purposes

8.4.3 Does Compression Level Differ in Chat and Chunk?

The compression measures described above were calculated for chat and chunk segments.

Compresssion for speech only was then measured on a chunk by chunk and chat by chat basis, and the distributions of the individual compression scores in chat and chunk phases were compared. It was found that compression in chunks ranged from 1.000 to 1.675, with mean of 1.084 (median - 1.062, IQR - 1.031 to 1.107) , while in chat the range was 1.000 to 1.605 with mean of 1.163 (median - 1.145, IQR - 1.081 to 1.235). Wilcox Rank Sum tests showed that compression levels differed significantly between chat and chunk ($W = 55373$, p-value < 2.2e-16), and were significantly higher in chat ($W = 55373$, p-value < 2.2e-16).

	Overall	A	B	C	D	E	F
DurPSL(O)	7937	1796	1146	1188	688	978	2142
DurFSL(O)	6371	1314	1025	1044	598	731	1660
ComSL(O)	**1.25**	**1.37**	**1.12**	**1.14**	**1.15**	**1.34**	**1.30**
DurPSL(X)	13058	2385	2676	2496	2188	1057	2255
DurFSL(X)	11728	2037	2537	2336	1974	887	1957
ComSL(X)	**1.11**	**1.17**	**1.05**	**1.07**	**1.11**	**1.19**	**1.15**
DurPS(O)	7187	1461	1086	1127	665	886	1962
DurFS(O)	6114	1224	998	1016	586	696	1594
ComS(O)	**1.18**	**1.19**	**1.09**	**1.09**	**1.13**	**1.27**	**1.23**
DurPS(X)	12422	2184	2595	2411	2099	1004	2129
DurFS(X)	11514	1979	2506	2299	1950	867	1913
ComS(X)	**1.08**	**1.10**	**1.04**	**1.05**	**1.08**	**1.16**	**1.11**
DurPL(O)	749	336	60	61	23	91	179
DurFL(O)	563	226	55	49	22	74	123
ComL(O)	**1.33**	**1.49**	**1.09**	**1.24**	**1.05**	**1.23**	**1.32**
DurPL(X)	636	201	81	85	89	53	127
DurFL(X)	518	146	73	68	85	47	100
ComL(X)	**1.23**	**1.38**	**1.11**	**1.25**	**1.05**	**1.13**	**1.27**
No. Sp	–	5	3	4	3	4	5

Table 8.4 – Compression of speech and laughter (ComSL), speech only (ComS), and laughter only (ComL) in dataset by phase. Compression is calculated by phase (chat=O,chunk=X) overall and per conversation, using durations of participant contributions for speech and laughter (DurPSL), speech only (DurPS) and laughter only (DurPL) and floor durations for speech and laughter (DurFSL), speech only (DurFS), and laughter only (DurFL). The number of speakers (No. Sp) in each conversation is given for reference purposes.

8.5 Floor State Changes around Silence and Overlap

This section focuses on *how* floor state changes happen in casual conversation, exploring how and when speakers take and leave the floor, in order to better understand turn change and retention.

The analysis uses the floor state annotation described in Chapter 4, Section 4.5. Only speech and silence were considered, with laughter disregarded and treated as silence, allowing the individual floor state labels and also strings of consecutive floor state labels to be simplified. To facilitate data processing, the GS label was changed to GX. Figure 8.4 illustrates the simplified labelling system used in the analyses de-

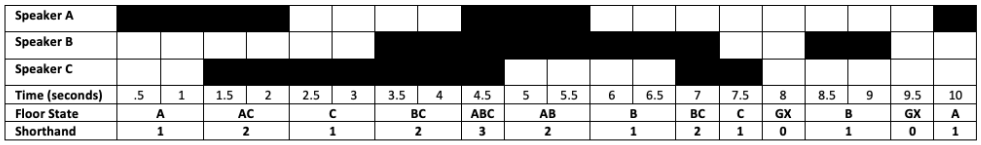


Figure 8.4 – Three participants in a 10-second stretch of conversation, where black represents speech and white represents silence. Speaker A produces 3 inter-pausal units (IPUs) totalling 4 seconds (2+1.5+.5) of speech, Speaker B produces 2 IPUs totalling 5 seconds (4+1), and Speaker C speaks for a total of 4.5 seconds(3.5+1) across 2 IPUs. There is global silence for 1 second (.5+.5). There are a total of 13 floor state labels. The entire floor state sequence would be represented as **A_AC_C_BC_ABC_AB_B_BC_C_GX_B_GX_A**, while the simpler numerical shorthand version is **1_2_1_2_3_2_1_2_1_0_1_0_1**. Note that interval durations are for illustration purposes and are somewhat unrealistic as overlap is generally shorter than in the example.

scribed below. As speech and silence are the phenomena of interest, and all of the conversation is taken up by speech (or overlap) and silence, the labels can be simplified to an initial representing the speaker(s) speaking in a particular interval and GX for global silence, resulting in a label set of 2^n possible labels. As the speaker identity is not considered in the analyses, the labels can be further simplified to integers representing the number of speakers speaking in a particular interval, giving a labelset of 0 for global silence, and 1 to n. Individual floor state labels can then be concatenated into sequences representing the activity over any stretch of conversation.

8.5.1 Floor State Transitions

To further understand how casual conversation progresses from one speaker to another, the following studies are based on conditions after an interval of single-party speech in the clear; an interval of speech by a single speaker which is not overlapped by any other speaker. [2]

1Sp is defined as such a stretch of single party speech, ***1Sp1*** as such an interval of duration one second or more, and ***1SpAny*** is used in contrast to denote such an

[2]Note that this interval is a floor state interval and may not comprise a complete interpausal unit (IPU).

interval of *any* duration greater than zero. In dyadic interaction, an interval of ***1Sp*** can theoretically end in silence, two-party overlap, or a *smooth switch* or transition to the other speaker. Here, smooth switch means a change from speaker A to B where the endpoint of A's utterance coincides with the start of B's (within a particular tolerance). In manually segmented data, the likelihood of such a smooth switch is low, given the precision of annotation software based on graphical interfaces (using Praat software results in possible error of 50ms when segmenting 10 seconds of audio on fullscreen display on a 23" monitor). As a result, smooth switches (e.g. **1_1** in the shorthand labels) and cases where two speakers start or stop speaking 'simultaneously' (e.g. **0_2** or **2_0**) should be very rare in the data.

The type of transition is further determined by the right hand context. To describe transition types, this study adapts the terminology used in (Heldner and Edlund, 2010) for dyadic interaction. In two-party interaction, silence can transition to ***1Sp*** resulting in a within or between speaker silence (1_0_1), while overlap can transition back to ***1Sp***, resulting in within or between speaker overlap (1_2_1). For speakers A and B, within speaker silence (WSS) is defined as **A_GX_A** and between speaker silence (BSS) is defined as **A_GX_B** where GX denotes global silence, while within and between speaker overlap are **A_AB_A** and **A_AB_B**. Thus, ***1Sp*** can transition back to ***1Sp*** with one intervening interval of silence or overlap, e.g. **1_0_1** or **1_2_1**.

For multiparty interaction, more possibilities emerge. Given the low likelihood of smooth switching and simultaneous onset or offset of speech, the more likely possibilities should be transitions between silence and ***1Sp*** by any participant, and between two-party overlap and ***1Sp*** by any speaker, as in dyadic interaction, and also transition to overlap involving more than two parties, e.g. **1_2_3**. Thus, in practice, multi-party speech should be able to transition back to ***1Sp*** with one intervening interval of silence or overlap, as in the two-party condition, but the number of intervening intervals would increase once 3- or more party overlap occurs.

Bigram	Overall (%)	Chat (%)	Chunk (%)
0_1	35.54	31.91	38.06
1_0	35.56	31.76	38.21
1_2	12.81	15.81	10.71
2_1	12.82	15.94	10.64
2_3	1.23	1.75	0.87
3_2	1.24	1.78	0.86
3_4	0.06	0.11	0.12
4_3	0.06	0.03	0.03

Table 8.5 – Eight most frequent bigram transitions (by % frequency) overall, for chat and chunk interaction. These bigrams account for 99.22% of all transitions in dataset.

While any number of intervals could, in theory, appear between two intervals of single-party speech in multiparty conversation, odd-numbers of intervals are expected to be more likely, e.g **1_2_3_2_3_4_3_2_1**. Odd numbers would allow return from overlap conditions entered into, while even numbers of intervals would entail unlikely smooth switch or simultaneous onset or offset of speech by two or more speakers. The above discussion is based on transitions without regard to the duration of the intervals of ***1Sp*** concerned.

8.5.2 General Floor State Bigrams

To investigate how floor transitions progress in casual talk, bigrams of the number of speakers in every two contiguous intervals in the dataset were created. As an example, **1_0** represented a transition from ***1Sp*** to silence while **3_2** represents a transition from three speakers to two speakers.

For five speakers, there are 34 bigram transitions possible in theory, if simultaneous starts and stops and smooth switches are permitted (0_0 and 5_5 are impossible, so 6x6-2)). Of these, 19 appeared in the data – 0_1, 0_2, 1_0, 1_1, 1_2, 1_3, 2_0, 2_1, 2_2, 2_3, 2_4, 3_1, 3_2, 3_3, 3_4, 4_2 ,4_3, 4_5, and 5_4. The 12 transitions reflecting simultaneous starts/stops by three or more participants did not (0_3, 0_4, 0_5, 1_4, 1_5, 2_5, 3_0, 4_0,5_0, 4_1,5_1, 5_2), nor did 3_5, 4_4, or 5_3.

Eight bigrams reflecting transitions adding or subtracting one speaker accounted

for 99.22% of the data, as shown in Table 8.5. The other 13 which also appeared, accounted for between 0.001% and 0.03% of transitions each.

8.5.3 *1Sp-* and *1Sp1-* Bigrams

Across the dataset, there were 14938 intervals of ***1Sp***. The proportions of transitions *to* silence and overlap *from* ***1Sp*** differed between chat and chunk. Cases of ***1Sp*** followed by silence (**1_0**) accounted for 66.30% of all transitions from ***1Sp*** in chat and 77.64% in chunk, while ***1Sp*** followed by two-party overlap (1_2) accounted for 33% of cases in chat and 21.77% in chunk. The proportions above apply to all intervals of ***1Sp***, regardless of duration. To more closely approximate conditions at the end of a turn, a lower bound of one second is placed on the duration of the initial ***1Sp*** interval and transitions are examined from this starting point. There were 5572 single-speaker intervals of 1 second or more (***1Sp1***). Smooth switches and simultaneous starts/stops by two participants accounted for 0.48% of data. Cases of ***1Sp1*** followed by silence (**1_0**) accounted for 75.07% of cases in chat and 83.99% in chunk, while ***1Sp1*** followed by two-party overlap accounted for 24.24% of cases in chat and 15.61% in chunk. In all bigram cases, transition from one-party speech to silence (1_0) is more common than transition from one-party speech to overlap (1_2). In all cases (general, ***1Sp***, ***1Sp1***), the proportion of transition from one-party speech to silence (1_0) is higher in chunk than in chat, while that of transition to overlap is higher in chat than in chunk (all significant at p<.001). Floor state transitions between single speakers are examined in the next section.

8.6 Transitions between Single Speakers

To investigate transitions between intervals of single party speech, a situation analagous to a turn change or retention, analysis was carried out of transitions from intervals of single party speech of at least one second's duration (***1Sp1***) to the next ***1Sp1*** interval

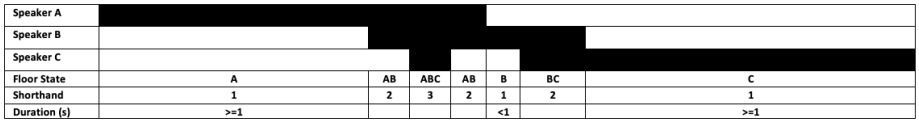


Figure 8.5 – Three participants in a stretch of conversation, where black represents speech and white represents silence. If we start at the left hand single speaker interval (A), which is of duration one second or longer, and find the next single speaker interval of any duration, we can label the sequence from A to B as a between speaker transition, **A_AB_ABC_AB_B** (**1_2_3_2_1**). However, if a one second threshold is also applied to the second single party speech interval in a transition, this would not be a valid transition as B is shorter than one second. With a right hand threshold applied the sequence from A to C would be a valid between speaker transition, **A_AB_ABC_AB_B_BC_C** (**1_2_3_2_1_2_1**).

of at least one second duration. This one-second minimum duration was chosen as a reasonable minimum length for an utterance. For these transitions, the possibilities multiply, as the intervening intervals can include ***1Sp*** segments shorter than the threshold duration, and thus sequences such as **1_2_1_0_1** are possible.

For convenience, we also define ***1SpAny*** as such an interval of *any* duration. For completeness, transitions from ***1Sp1*** to ***1SpAny*** were checked and compared with transitions (***1Sp1***) to the next ***1Sp1***. These transitions can be operationalized by placing a lower bound on the first and final single speaker interval durations (the left and right hand boundaries).

For this work, we retain Heldner et al's notion of within and between speaker phenomena, but, as multiparty transitions can involve a combination of overlap and silence, we define only two transition types – *within speaker transitions* (WST) where we examine transitions beginning and ending with the same speaker, and *between speaker transitions* (BST), which start with one single speaker and transition to another single speaker.

Figure 8.5 demonstrates how annotation of a within or between speaker transition is affected by the choice of threshold duration for the right hand interval in the transition. If a one second threshold is applied to the opening or left hand single speaker

interval, and any duration is allowed for the right hand or closing interval of single party speech the diagram would show a between speaker transition from A to B, with 3 intervening intervals, whereas if a threshold of one second is applied to both the opening and closing intervals, the diagram would show a between speaker transition from A to C with 5 intervening intervals.

8.6.1 *1Sp1-1Sp1* Transitions

For each ***1Sp1*** interval, a search was performed forward in the dataset to locate the next ***1Sp1***. The sequence of intervals (in terms of speaker numbers) from the initial ***1Sp1*** to the next ***1Sp1*** were extracted. These sequences could be of any length, and are thus termed transition interval n-grams. As an example, **1_2_3_2_1_0_1** represents a sequence where ***1Sp1*** transitions to 2-, 3-, and back to 2-party overlap before returning to ***1Sp*** (<1sec) by the original speaker or another, and then transitions to silence before finally transitioning to ***1Sp1*** by the original speaker or another. This example contains 5 intervening intervals between the two stretches of ***1Sp1***. It can be seen that such cases can contain both overlap and silence, and are termed between speaker transition (BST) or within speaker transition (WST) depending on the identities of the first and last speakers - if the stretch begins and ends with the same speaker, it is termed a WST, while a stretch beginning and ending with different single speakers is a BST.

Table 8.6 shows the distributions of ***1Sp1***-***1Sp1*** transitions overall, in chat and in chunks. Overall, 95.53% of all ***1Sp1-*** intervals are closed by a later ***1Sp1*** in fewer than 16 intervening intervals. For the remaining 4.47% of ***1Sp1-*** intervals, labelled 16+, at least 16 intervals occurred before a second ***1Sp1*** interval was encountered. Even-number cases accounted for only 112 (2.1%) of the 5382 transitions between 1 and 15 intervals long. Overall, the most frequent class of transitions are those with one intervening interval which account for 41.13% of cases.

Intervals	Overall (%)	Chat (%)	Chunk (%)
1	41.13	30.95	45.78
2	0.50	0.51	0.49
3	21.76	19.83	22.64
4	0.39	0.46	0.37
5	12.66	14.44	11.84
6	0.27	0.34	0.23
7	7.67	9.17	6.98
8	0.22	0.34	0.15
9	4.45	6.25	3.63
10	0.27	0.46	0.18
11	2.98	4.06	2.48
12	0.16	0.29	0.10
13	2.01	3.15	1.49
14	0.19	0.34	0.13
15	0.86	1.09	0.76
16+	4.47	8.31	2.72

Table 8.6 – Number of intervening intervals per ***1Sp1-1Sp1*** transition (by % frequency) overall, and for chat and chunk interaction.

Distributions of transitions between single-speaker intervals of 1 second or more in duration in chat and chunk are shown in Figure 8.6, where it can be seen that the vast majority of intervening intervals are in stretches of odd numbers of intervals. Comparing chat and chunk conditions, 30.94% of transitions in chat occurred with one intervening interval, compared to 45.78% in chunk.

In both chat and chunk, disregarding the even-number cases, the number of transitions declines monotonically with the number of intervening intervals between ***1Sp1*** intervals. The chat condition starts with a smaller percentage of 1-interval transitions and declines at a lower rate than the chunk condition.

In both conditions, it is likely that numbers continue to decline with increasing intervals in a long tail. The 16+ category, a bucket category, is more than three times as large proportionally in chat (8.31%) as in chunk(2.72%).

8.6.2 Between and Within Speaker Transitions

In terms of ***1Sp1-1Sp1*** overall transitions, 21.74% were BSTs while WSTs accounted for 73.78%. The odd numbered cases and the 16+ interval bucket class were excluded

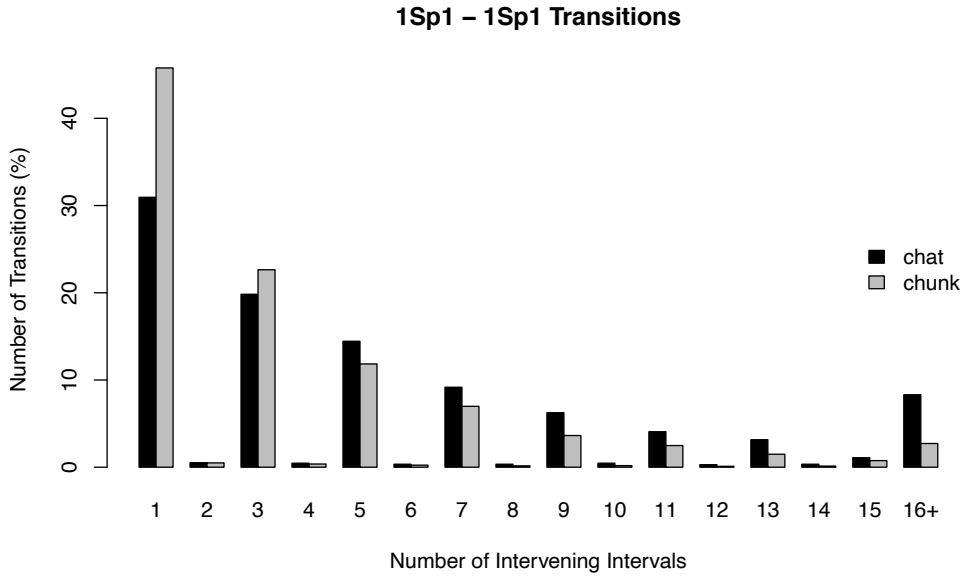


Figure 8.6 – Number of intervening floor state intervals between transitions beginning and ending with single-speaker intervals of 1 second or more in duration (textbf1Sp1-1Sp1)

from the ***1Sp1-1Sp1*** transition data, leaving 5270 transitions, comprising 77.24% WST and 22.76% BST with intervening intervals ranging from 1 to 15. Figure 8.7 shows these BST and WST transitions by number of intervals, while Figure 8.8 shows interval types in chat and chunk phases, and the proportion of transitions per interval total. One-interval transitions were the largest group for BST and WST for both chat and chunk, with the proportion of 1-interval transitions particularly high for WST, and very much so in the case of chunk

The distributions of transition n-grams between intervals of one speaker speaking in the clear for at least one second (***1Sp1-1Sp1***) show that speaker transitions in chat are spread over more intervening intervals than in chunk, thus increasing the frequency of more complex transitions. This reflects more activity in chat. Within speaker transitions are more likely to be one-interval, perhaps reflecting breathing pauses in a single speaker, particularly so in chunk. It would be very interesting to separate within speaker breathing pauses from other transitions in order to bet-

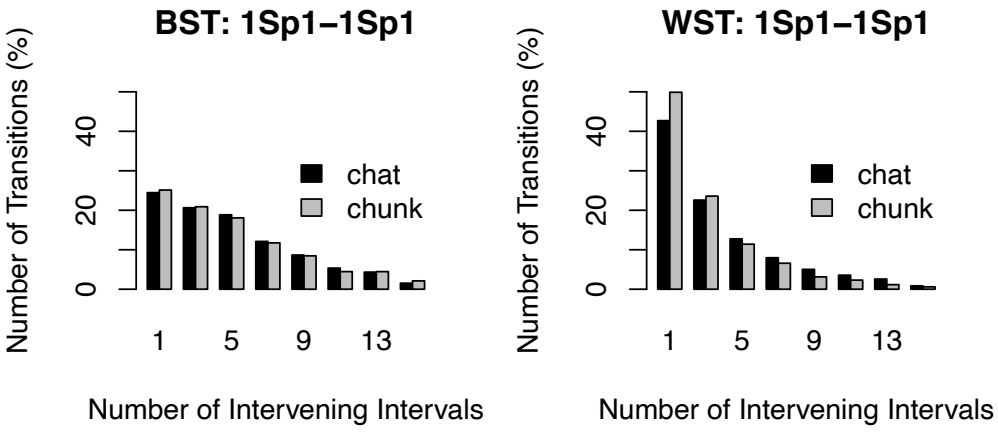


Figure 8.7 – Number of floor state intervals between single-speaker intervals of 1 second or more in duration (***1Sp1***) in Between Speaker Transitions (BST, left) and Within Speaker Transitions (WST, right) of between 1 and 15 intervals in chat and chunk phases.

ter understand silence around turn change and retention. One-interval transitions comprise the largest class, with a higher proportion of one-interval transitions in chunk than chat, and higher proportions of within speaker than between speaker one-interval transitions in both conditions, particularly in chunk. However, one-interval transitions only account for 41.03% of transitions overall, reflecting the need to consider more complex transitions around turn change and retention.

8.6.3 *1Sp1-1Sp1* vs *1Sp1-1SpAny* Transitions

In order to test how the duration of the right hand single speaker interval affected the distribution of transitions, transition n-grams were generated from each initial single speaker interval (***1Sp1***) to the next single speaker interval *of any duration* (***1SpAny***), as shown in Figure 8.9. It can be seen that the vast bulk of the ***1Sp1-1SpAny*** transitions occur with one intervening interval (95.97%), while 3, 5, and 7 intervals account for 3.01%, 0.57%, and 0.16% overall respectively.

Comparing ***1Sp1-1SpAny*** transitions sharing their left-hand ***1Sp1*** intervals with the 5270 ***1Sp1-1SpAny*** transitions treated in Section 8.6.2 results in the confusion ma-

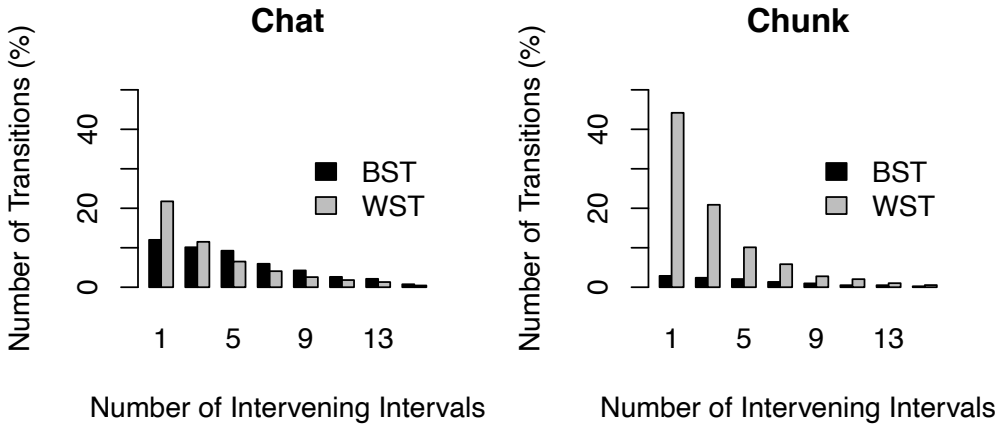


Figure 8.8 – Percentage of Between and Within Speaker Transitions per number of floor state intervals between single-speaker intervals of 1 second or more (***1Sp1-1Sp1***) in chat and chunk phases.

		1Sp1-1Sp1		
		BST	**WST**	**Sum**
	BST	15.81	14.21	30.01
1Sp1-1SpAny	**WST**	6.95	63.03	69.99
	Sum	22.76	77.24	100

Table 8.7 – Confusion Matrix (%)

trix in Table 8.7.

The transition type label changes for 21.15% of transitions overall depending on how the right hand interval is defined, while 56% of transitions would change the number of intervening intervals involved. For chat, a larger proportion of labels (27.87%) change (13.56% Between and 14.31% Within), as chat contains a higher proportion of intervening intervals involving more than one label. In chunk, the proportion is smaller (18.28% - 14.48% Between, 3.80% Within), largely because the bulk of transitions in chunk are BST, which are largely one interval.

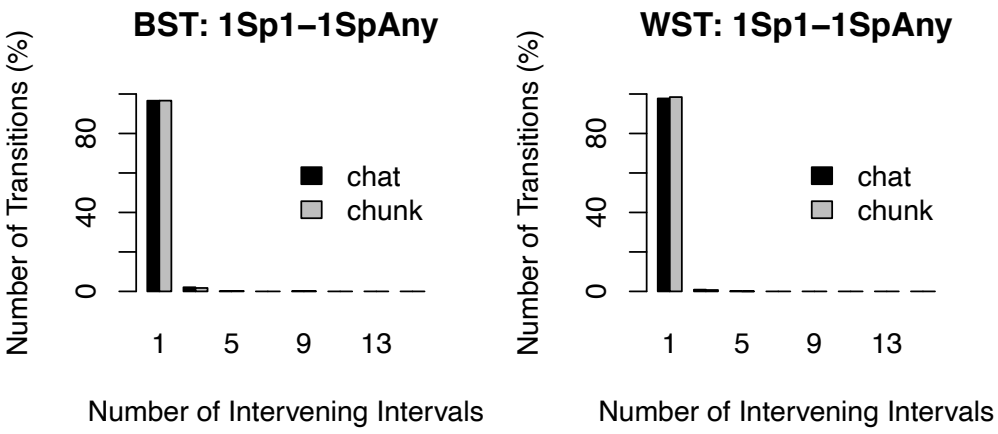


Figure 8.9 – Number of floor state intervals between (***1Sp1***) and (***1SpAny***) in Between Speaker Transitions (BST, left) and Within Speaker Transitions (WST, right).

8.7 Conclusions

This chapter has examined local changes in floor state, liveliness, duration of individual contributions and of overlap and silence, and prosodic conditions around silence and overlap. The results are discussed in Chapter 10.

Chapter 9

Comparison With Task-Based Conversation

This chapter describes experiments comparing chat and chunk segments of the CasualTalk dataset with TEAMS, a corpus of collaborative games played by three or four party teams. This comparison aimed to explore differences in the distribution of silence, speech and overlap in the two corpora, and to compare characteristics of chat and chunk phases with task-based talk. This chapter addresses Research Question 3

> **How do casual talk and chat and chunk phases compare to task-based dialogue in terms of speech, silence, overlap, liveliness, and conditions around floor state changes?**

9.1 Working Dataset and Pre-processing

This section describes the preparation performed to compare the TEAMS Corpus of task-based collaborative games, described in Chapter 3, Section 3.8, with CasualTalk as a whole, and with the chat and chunk phases in CasualTalk.

9.1.1 The TEAMS Corpus

The TEAMS Corpus (Litman et al., 2016) comprises 47 hours of multiparty interaction from 62 teams (35 three-person and 27 four-person), playing a collaborative board game – Forbidden Island. This game requires cooperation and communication among the players to win as a group. The games were played in two conditions, Control and Intervention. In both conditions the teams played the game twice. In the Intervention condition players were also given ten minutes of training in co-operation strategy and were allowed extra time between Game 1 and Game 2 to discuss how their performance could be improved. For the purposes of this analysis, TEAMS recordings of Game 1 for the Control condition were used. There were a total of 31 Game 1 interactions in the Control condition, comprising 20 3-person and 11 4-person games. There were 104 participants in these games, 65 female and 39 male, with ages ranging from 18 to 61 (mean = 25.03, sd =10.99). The 31 games comprised a total of 14 hours of conversation time.

To compare with the CasualTalk dataset, the existing TEAMS segmentations and annotations for Game 1 in the control condition were processed to generate a set of speech and silence annotations. Laughter was treated as silence. Praat TextGrids were generated for each game with reference to the original wav sound files. To explore the dynamics of the different speech-exchange systems, the recordings and transcripts for each game were processed using Praat to create 'floor state' annotations, as described for the CasualTalk dataset in Chapter 4, Section 4.5. These annotations divided each interaction into labelled intervals, where an alphanumeric code for each interval recorded who was speaking during the interval timespan, or labelled intervals of global silence (where nobody was speaking). For example, the label **aSbS** denotes that speakers **a** and **b** are speaking in overlap, while **cS** indicates that **c** is speaking alone, and **GX** denotes global silence. Any speakers not mentioned in label are silent for the interval described.

The resulting annotations were converted to an R dataframe and statistics for speech, silence, overlap and speaker change dynamics in the two datasets were compared using R statistical software (R Core Team, 2014). The games were compared to chat and chunk phases of larger casual conversations as these separate phases constitute speech exchange systems or genres in their own right (Bakhtin, 1986). The resulting dataset (henceforth TEAMS1), at 14 hours of conversation time is considerably larger than CasualTalk, at 6.5 hours, a reflection of the scarcity of casual talk data.

9.2 Conversational Floor Distribution

The TEAMS1 and CasualTalk data were combined and overall statistics for the proportion of interaction in casual talk and games are shown in Table 9.1

Number of Speakers	Casual	Game
0 (s)	5396.97	15451.22
0 (%)	23.44	30.59
1 (s)	15765.87	26719.72
1 (%)	68.47	52.89
2 (s)	1750.06	7352.74
2 (%)	7.60	14.55
3+ (s)	112.67	991.81
3+ (%)	0.49	1.96

Table 9.1 – Conversation time in seconds and as a percentage of total conversation time taken up by 0, 1 ,2 ,3+ speakers in Game and Casual Talk

There was less global silence in casual talk than in games (23.44% vs. 30.59%), considerably more one-party speech (68.4% vs. 52.89%), and less overlap (8.1% vs. 16.5%).

For a more detailed view of distribution of silence, one-party speech and overlap, Table 9.2 contrasts the proportion of the interaction by 0, 1, or 3+ speakers in game, chat phases of conversation, and chunk phases of conversation. Figure 9.1 shows

Number of Speakers	Game	Chat	Chunk
0 (s)	15451.22	2122.92	3274.05
0 (%)	30.59	25.78	22.14
1 (s)	26719.72	5071.43	10694.45
1 (%)	52.89	61.58	72.31
2 (s)	7352.74	977.01	733.04
2 (%)	14.55	11.86	5.23
3+ (s)	991.81	64.64	48.03
3+ (%)	1.96	0.78	0.32

Table 9.2 – Conversation time in seconds and as a percentage of total conversation time taken up by 0, 1 ,2 ,3+ speakers in Game and Casual Talk Chat and Chunk phases

the proportion of the interaction occupied by 0, 1, or 2+ speakers in chat and chunk phases and in game. Overlap involving three or more speakers is much less frequent than two party overlap in all three conditions, and thus 2 and 3+ party overlap were amalgamated to 2+ overlap for clarity on the figure.

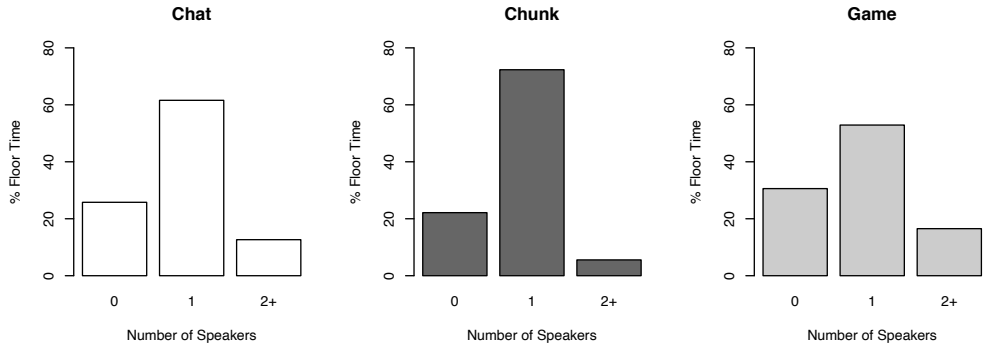


Figure 9.1 – Distribution of the floor in chat (white), chunk (dark grey), and games (light grey). X-axis shows number of speakers (0,1,2+) speaking concurrently.

The boxplots in Figure 9.2 show the distributions of the proportions of silence, one-party speech, and overlap per individual chat, chunk, or game.

In all conditions the most common conversational situation in terms of time taken was single participant speaking in the clear, the second most common was global silence, with overlap accounting for much less of the conversational time. Non-parametric

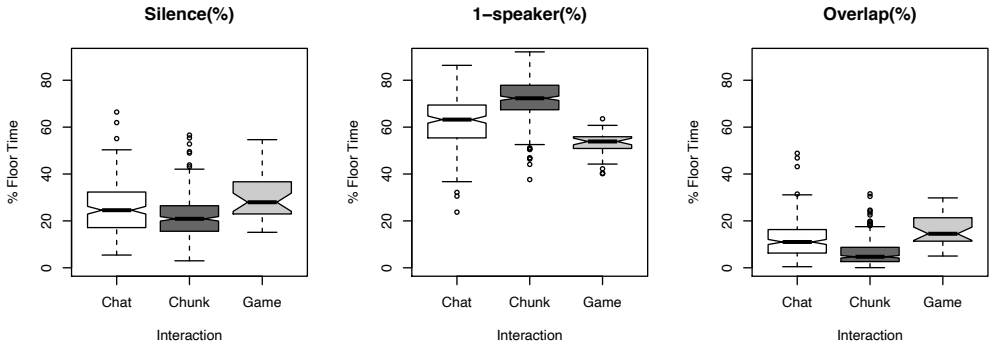


Figure 9.2 – Distribution of the floor in Silence, 1-Speaker and Overlap in chat (white), chunk (dark grey), and games (light grey).

Wilcoxon Rank Sum tests were applied to examine whether proportions of silence, single speaker, and overlap differed between the Chat, Chunk, and Game conditions.

Wilcoxon Rank Sum tests showed significant differences in the proportion of silence in Game and Chat (W = 9242, p-value = 1.689e-06), and in the proportion of silence in Chat and Chunk (W = 27449, p-value = 3.148e-05), with Game interactions having significantly more silence (median 27.%) than Chat (median 24.6%), W = 9242, p-value = 8.444e-07. and that Chat segments in turn having significantly more silence than Chunk segments (median 20.9%), W = 27449, p-value = 1.574e-05. Wilcoxon Rank Sum Tests showed significant differences in the proportion of one-party speech in Chunk and Chat (W = 54268, p-value < 2.2e-16) and in Chat and Game (W = 2306, p-value = 6.396e-15), and that there was significantly more 1-party speech in Chunk (median 72.3%) than in Chat (median 63.2%), W = 54268, p-value < 2.2e-16, and in Chat than in Game conditions (median 53.9%), W = 2306, p-value = 3.198e-15.

Wilcoxon Rank Sum Tests showed showed significant differences in the proportion of overlap differed significantly in Game and in Chat (W = 8270, p-value = 0.002496) and in CHat and Chunk (W = 17261, p-value < 2.2e-16), and that overlap was significantly higher in Game (median 14.5%) than in Chat (median 11%),W = 8270, p-value = 0.001248, and in Chat than in Chunk (4.7%), W = 17261, p-value < 2.2e-16.

9.3 Speaker Change Activity

The TEAMS1 dataset contains 64015 changes in speech/silence configuration, an average floor change rate of 1.27 per second. The CasualTalk dataset contains 30688 changes, an average floor change rate of 1.33 per second. Chunks account for 18081 of these changes, an average 1.22 per second, while there are 12607 changes in the chat data, for an average 1.5 per second floor state change rate. The average compression (speech only) for TEAMS1 was 1.26, compared to 1.084 for chunks, and 1.163 for chat.

9.3.1 Floor State Transitions

The following analyses are based on conditions after an interval of single-party speech in the clear; an interval of speech by a single speaker which is not overlapped by any other speaker. [1] The methods described in Chapter 4, Section 4.5, are used here to further compare the dynamics of causal conversation's chat and chunk phases with the TEAMS games. A brief recap of the methods is given below.

As described in Chapter 8, Section 8.5.1, ***1Sp*** is defined as a stretch of single party speech in the clear, ***1Sp1*** as such an interval of duration one second or more, and ***1SpAny*** is used in contrast to denote such an interval of *any* duration greater than zero. The type of transition is further determined by the right hand context. In two-party interaction, silence can transition to ***1Sp*** resulting in a within or between speaker silence (1_0_1), while overlap can transition back to ***1Sp***, resulting in within or between speaker overlap (1_2_1). For speakers A and B, within speaker silence (WSS) is defined as **A_GX_A** and between speaker silence (BSS) is defined as **A_GX_B** where GX denotes global silence, while within and between speaker overlap are **A_AB_A** and **A_AB_B**. Thus, ***1Sp*** can transition back to ***1Sp*** with one intervening interval of silence or overlap, e.g. **1_0_1** or **1_2_1**.

[1] Reminder:Note that this interval is a floor state interval and may not comprise a complete inter-pausal unit (IPU).

Bigram	Game (%)	Chat (%)	Chunk (%)
0_1	27.48	31.91	38.06
1_0	27.49	31.76	38.21
1_2	18.53	15.81	10.71
2_1	18.55	15.94	10.64
2_3	3.66	1.75	0.87
3_2	3.67	1.78	0.86
3_4	0.21	0.11	0.03
4_3	0.22	0.12	0.03

Table 9.3 – Eight most frequent bigram transitions (by % frequency) in game, chat and chunk interaction. These bigrams account for 99.56% of all transitions in dataset.

For multiparty interaction, more possibilities emerge. Given the low likelihood of smooth switching and simultaneous onset or offset of speech, the more likely possibilities should be transitions between silence and ***1Sp*** by any participant, and between two-party overlap and ***1Sp*** by any speaker, as in dyadic interaction, and also transition to overlap involving more than two parties, e.g. **1_2_3**. While any number of intervals could, in theory, appear between two intervals of single-party speech in multiparty conversation, odd-numbers of intervals are expected to be more likely, e.g **1_2_3_2_3_4_3_2_1**. Odd numbers would allow return from overlap conditions entered into, while even numbers of intervals would entail unlikely smooth switch or simultaneous onset or offset of speech by two or more speakers. The above discussion is based on transitions without regard to the duration of the intervals of ***1Sp*** concerned.

9.3.2 General Floor State Bigrams

To compare how floor transitions progress in chat, chunk, and game speech exchange systems, bigrams of the number of speakers in every two contiguous intervals in the dataset were created. As an example, **1_0** represented a transition from ***1Sp*** to silence while **3_2** represents a transition from three speakers to two speakers.

Eight bigrams reflecting transitions adding or subtracting one speaker accounted for 99.56% of the data, as shown in Table 9.3.

9.3.3 *1Sp*- and *1Sp1*- Bigrams

Across the dataset (TEAMS1 plus CasualTalk), there were 75850 intervals of ***1Sp***.

The proportions of transitions *to* silence and overlap *from* ***1Sp*** differed between chat and chunk. Cases of ***1Sp*** followed by silence (**1_0**) accounted for 66.30% of all transitions from ***1Sp*** in chat and 77.64% in chunk and for 59.63% of transitions in game, while ***1Sp*** followed by two-party overlap (1_2) accounted for 33% of cases in chat, 21.77% in chunk, and 40.19% in game. The proportions above apply to all intervals of ***1Sp***, regardless of duration.

To more closely approximate conditions at the end of a turn, a lower bound of one second is placed on the duration of the initial ***1Sp*** interval and transitions are examined from this starting point. There were 24726 single-speaker intervals of 1 second or more (***1Sp1***). Smooth switches and simultaneous starts/stops by two participants accounted for 0.24% of these transitions. Cases of ***1Sp1*** followed by silence (**1_0**) accounted for 75.07% of cases in chat, 83.99% in chunk and 72.31% in game, while ***1Sp1*** followed by two-party overlap accounted for 24.24% of cases in chat, 15.61% in chunk and 27.52% in game. In all bigram cases, transition from one-party speech to silence (1_0) is more common than transition from one-party speech to overlap (1_2). In all cases (general, ***1Sp***, ***1Sp1***), the proportion of transition from one-party speech to silence (1_0) is highest in chunk than in chat with game having the lowest proportion, while that of transition to overlap is higher in game than in chat with chunk having the lowest proportion (all significant at $p<.001$).

Floor state transitions between single speakers are examined in the next section.

9.4 Transitions between Single Speakers

This section uses the methods described in Chapter 4, Section 4.5 to compare game, chat and chunk speech in terms of transitions between intervals of single party speech, a situation analagous to a turn change or retention, analysis was carried out of transi-

tions from intervals of single party speech of at least one second's duration (***1Sp1***) to the next ***1Sp1***. This one-second minimum duration was chosen with reference to the distributions of IPU duration in the data - where the median length of single speaker IPUs longer than 400ms (to exclude backchannels) was 1.2s. For convenience, we also define ***1SpAny*** as such an interval of *any* duration. For completeness, transitions from ***1Sp1*** to ***1SpAny*** were checked and compared with transitions (***1Sp1***) to the next ***1Sp1***.

9.4.1 *1Sp1-1Sp1* Transitions

For each ***1Sp1*** interval, a search was performed forward in the dataset to locate the next ***1Sp1***. The sequence of intervals (in terms of speaker numbers) from the initial ***1Sp1*** to the next ***1Sp1*** were extracted. These sequences could be of any length, and are thus termed transition interval n-grams. As an example, **1_2_3_2_1_0_1** represents a sequence where ***1Sp1*** transitions to 2-, 3-, and back to 2-party overlap before returning to ***1Sp*** (<1sec) by the original speaker or another, and then transitions to silence before finally transitioning to ***1Sp1*** by the original speaker or another. This example contains 5 intervening intervals between the two stretches of ***1Sp1***. It can be seen that such cases can contain both overlap and silence, and are termed between speaker transition (BST) or within speaker transition (WST) depending on the identities of the first and last speakers - if the stretch begins and ends with the same speaker, it is termed a WST, while a stretch beginning and ending with different single speakers is a BST.

Table 9.4 shows the distributions of ***1Sp1-1Sp1*** transitions in game, chat, and chunk interaction. Overall, 93.04% of all 24700 ***1Sp1-*** intervals are closed by a later ***1Sp1*** in fewer than 16 intervening intervals. For the remaining 4.47% of ***1Sp1-*** intervals, labelled 16+, at least 16 intervals occurred before a second ***1Sp1*** interval was encountered. Even-number cases accounted for only 245 (1.06%) of the transitions

Intervals	Game (%)	Chat (%)	Chunk (%)
1	35.05	30.95	45.78
2	0.08	0.51	0.49
3	18.97	19.83	22.64
4	0.13	0.46	0.37
5	12.54	14.44	11.84
6	0.11	0.34	0.23
7	8.87	9.17	6.98
8	0.13	0.34	0.15
9	6.18	6.25	3.63
10	0.09	0.46	0.18
11	4.42	4.06	2.48
12	0.09	0.29	0.10
13	3.24	3.15	1.49
14	0.08	0.34	0.13
15	2.35	1.09	0.76
16+	7.69	8.31	2.72

Table 9.4 – Number of intervening intervals per ***1Sp1-1Sp1*** transition (by % frequency) the three different interaction types – game, chat and chunk.

between 1 and 15 intervals long.

Overall, the most frequent class of transitions are those with one intervening interval which account for 36.42% of cases.

Distributions of transitions between single-speaker intervals of 1 second or more in duration in game, chat and chunk are shown in Figure 9.3, where it can be seen that the vast majority of intervening intervals in chat and chunk and game are in stretches of odd numbers of intervals. In game, chat and chunk, disregarding the even-number cases, the number of transitions declines monotonically with the number of intervening intervals between ***1Sp1*** intervals.

Both chat and game conditions start with a smaller percentage of 1-interval transitions and decline at a lower rate than the chunk condition. The 16+ bucket category is considerably larger in chat (8.31%) and in game (7.69%) than in chunk (2.71%).

9.4.2 Between and Within Speaker Transitions

The odd numbered cases and the 16+ interval bucket class were excluded from the ***1Sp1-1Sp1*** transition data, leaving 22735 transitions, comprising 58.32% WST and

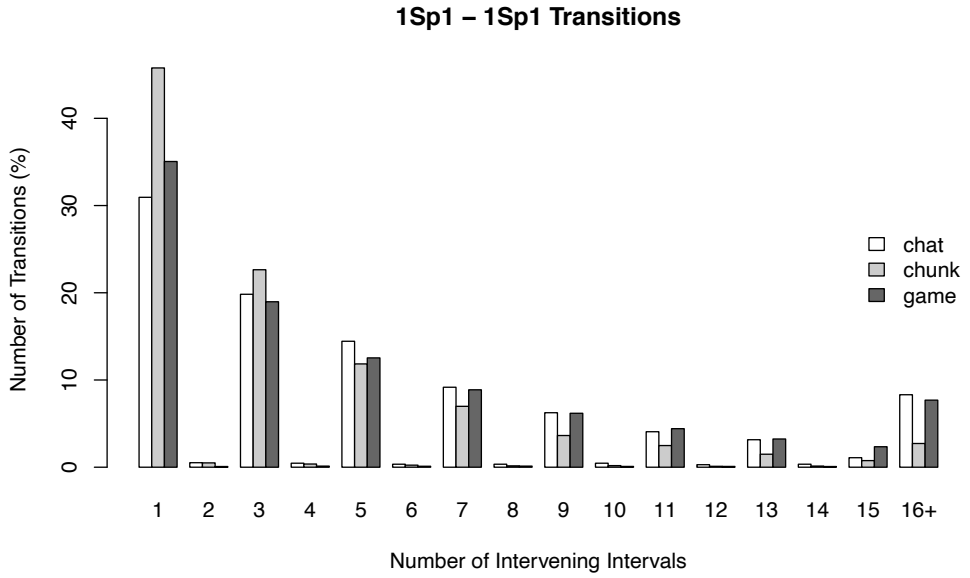


Figure 9.3 – Number of floor state intervals between single-speaker intervals of 1 second or more in duration

41.68% BST with intervening intervals ranging from 1 to 15. Figure 9.4 shows these BST and WST transitions by number of intervals, while Figure 9.5 shows interval types in chat and chunk phases, and the proportion of transitions per interval total. One-interval transitions were the largest group for BST and WST for all conditions, with the proportion of 1-interval transitions particularly high for WST, and very much so in the case of chunk.

The distributions of transition n-grams between intervals of one speaker speaking in the clear for at least one second (***1Sp1-1Sp1***) show that speaker transitions in chat and game are spread over more intervening intervals than in chunk, thus increasing the frequency of more complex transitions. This reflects more activity in chat and in game. Within speaker transitions are more likely to be one-interval, perhaps reflecting breathing pauses in a single speaker, particularly so in chunk. One-interval transitions comprise the largest class, with a higher proportion of one-interval transitions in chunk than in chat or game, and higher proportions of within speaker than be-

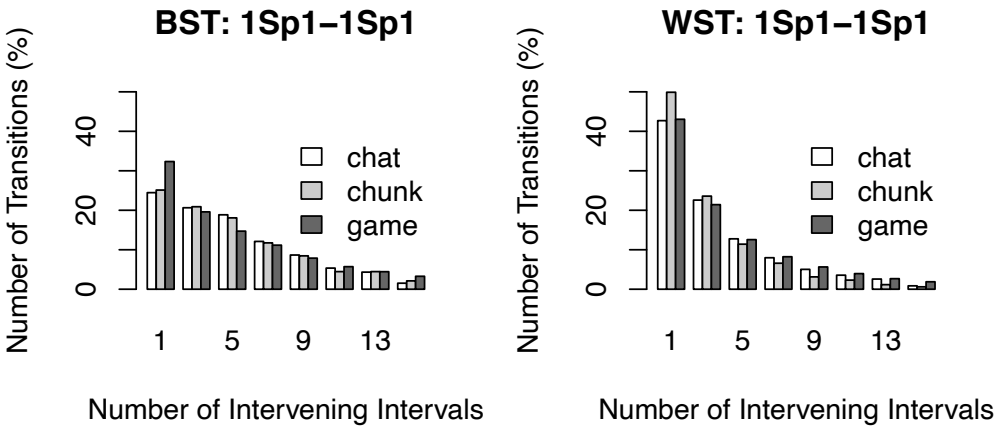


Figure 9.4 – Number of floor state intervals between single-speaker intervals of 1 second or more in duration (***1Sp1***) in Between Speaker Transitions (BST, left) and Within Speaker Transitions (WST, right) of between 1 and 15 intervals in chat, chunk and game.

tween speaker one-interval transitions in all conditions, particularly in chunk. However, one-interval transitions only account for 36.42% of transitions overall, reflecting the need to consider more complex transitions around turn change and retention.

9.4.3 *1Sp1-1Sp1* vs *1Sp1-1SpAny* Transitions

In order to test how the duration of the right hand single speaker interval affected the distribution of transitions, transition n-grams were generated from each initial single speaker interval (***1Sp1***) to the next single speaker interval *of any duration* (***1SpAny***), as shown in Figure 9.6. It can be seen that the vast bulk of the ***1Sp1-1SpAny*** transitions occur with one intervening interval (96.38%), while 3, 5, and 7 intervals account for 2.57%, 0.46%, and 0.09% overall respectively.

Comparing ***1Sp1-1SpAny*** transitions sharing their left-hand ***1Sp1*** intervals with the 22735 ***1Sp1-1Sp1*** transitions treated in Section 9.4.2 results in the confusion matrix in Table 9.5.

The transition type label changes for 21.15% of transitions overall depending on how the right hand interval is defined, while 59.96% of transitions would change the

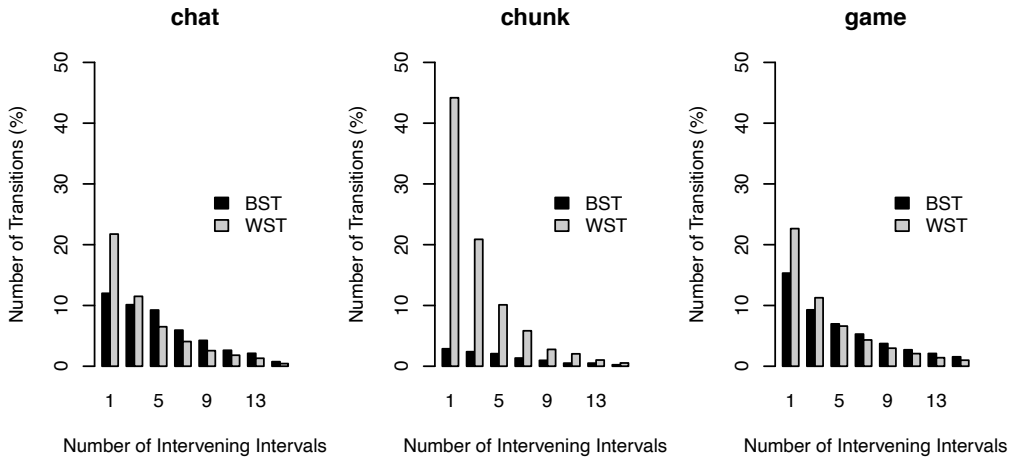


Figure 9.5 – Percentage of Between and Within Speaker Transitions per number of floor state intervals between single-speaker intervals of 1 second or more (***1Sp1-1Sp1***) in chat, chunk and game.

		1Sp1-1Sp1		
		BST	**WST**	**Sum**
	BST	31.87	16.23	48.11
1Sp1-1SpAny	**WST**	9.80	42.08	51.89
	Sum	41.68	58.32	100

Table 9.5 – Confusion Matrix (%)

number of intervening intervals involved. As seen in Chapter 8, Section 8.6.3 a larger proportion of labels (27.87% – 13.56% Between and 14.31% Within) change in chat than in chunk (18.28% - 14.48% Between, 3.80% Within). For the game condition, the proportions of changed labels is similar to those in chat, with 27.51% labels changing transition type, 16.84% Between and 10.67% Within Speaker transitions.

9.5 Conclusions

This chapter contrasted casual conversation in the CasualTalk dataset with TEAMS, a corpus of multiparty collaborative games. Speech, silence and overlap distributions were compared between casual talk in general and its chat and chunk phases with the games in TEAMS. Conditions around silence and overlap were also compared. Results

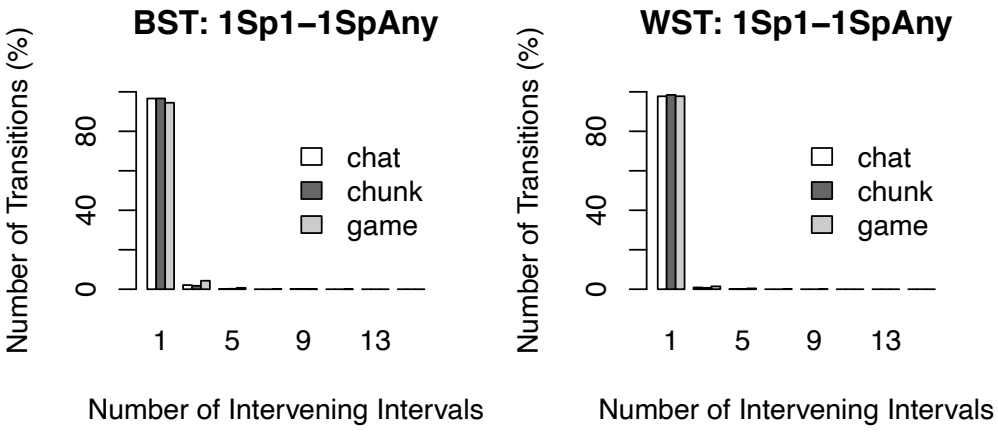


Figure 9.6 – Number of floor state intervals between (***1Sp1***) and (***1SpAny***) in Between Speaker Transitions (BST, left) and Within Speaker Transitions (WST, right).

are discussed in Chapter 10.

Part IV

Discussion and Conclusions

Chapter 10

Discussion of Results

In this chapter, the results from the experiments performed in preceding chapters are summarised and discussed.

The work described in this thesis is intended to further understanding of mulit-party casual social talk. The analysis of conversations covered the composition of casual talk and its constituent chat and chunk phases through analysis of its components: speech, silence, overlap, laughter and disfluency. The dynamics of casual talk, chat and chunk were analysed through examination of the floor state and its changes, the distribution of chat and chunk phases and their duration, liveliness, and prosodic conditions around silence and overlap. The CasualTalk dataset was compared with the TEAMS corpus of collaborative games to investigate differences between causal and task-based dialog in general, and particularly in interactions of chat, chunk, and game viewed as distinct speech exchange systems.

The work is based on 3 questions posed in Chapter 1.

What is the compostition of multiparty casual talk, and of its constituent chat and chunk phases in terms of measures of speech, silence, laughter and disfluency?

What are the characteristics of casual talk, and its constituent chat and

chunk phases, in terms of floor state dynamics at the utterance level?

How do casual talk and chat and chunk phases compare to task-based dialogue in terms of speech, silence, overlap, liveliness, and conditions around floor state changes?

These questions prompt exploration and description of casual talk and its chat and chunk phases, and of how chat and chunk phases can be differentiated in different ways, based on a range of low level features. One goal of the work is to add quantitative statistical description to the largely qualitative and theoretical descriptions in the literature describing chat and chunk structure. Another is to explore the composition of chat and chunk segments to see how they differ in terms of structure, timing, prosody, laughter, and disfluency. A third goal is to compare multiparty casual conversation with other multiparty speech systems. A final goal is to explore how insight into the structure of such talk impacts the requirements for dialog systems capable of social as well as task-based interaction. Below, the results obtained relevant to the Research Questions above and Research Outputs stated in Chapter 1 are summarized and discussed, with references to the chapters and sections in which relevant work was described.

10.1 Data and Annotation

An annotation scheme was created based on Slade and Eggins' definitions of chat and chunk phases (Chapter 4 Section 4.7). The scheme was tested by two annotators, resulting in a simple agreement of over 87% and a Cohen's Kappa of 0.74, indicating that the scheme created is robust. It was also found that annotating on the text transcripts alone was very effective for marking chat and chunk phases. This is a positive sign that annotation could eventually be performed on text transcripts alone, often a far more feasible solution than human annotation of audio or video data. There is also scope

for automatic tagging of chat and chunk segments in future work through development of stochastic models of chat and chunk sequences based on the annotations of the CausalTalk corpus and new annotations of other corpora using the annotation scheme.

A related research output is the creation of annotations. The annotated CasualTalk corpus comprises 6.5 hours of multiparty casual talk, and has been manually segmented, labelled for speech, silence, overlap and laughter, transcribed, and annotated for chat and chunk phases. The annotation scheme showed a high level of inter-rater reliability in use, indicating that the definitions are not subjective, but can be used to objectively identify and annotate chunk segments in other data. The process used to prepare disfluency (Chapter 4 Section 4.8) and floor state annotations (Chapter 4 Section 4.5) generated scripts and procedures which will be of use for similar tasks on other data.

10.2 Composition

10.2.1 Overall Composition of Casual Conversation

The overall composition of the conversations in CasualTalk were analysed. The overall proportion of the conversational time taken up by single party speech (68.49%), silence (23.43%) and overlap (8.08%) across the dataset (Chapter 6 Section 6.2.2) , can be compared to the proportions in TEAMS1, the subsection of the TEAMS corpus analysed in Chapter 9, and also in the ICSI corpus of multiparty meetings, comprising 75 meetings of between 3 and 9 participants totalling 67 hours of interaction, for which composition figures are available.

In ICSI conversational time was taken up by single party speech (66.9%), silence (27.8%) and overlap (5.28%). Thus, there is more slightly more conversational time devoted to speech in the CasualTalk dataset, but almost twice as much time devoted to overlap in casual talk as in the meetings corpus. In TEAMS1 the figures were single

party speech (52.9%), silence (30.59%) and overlap (16.51%), with much less single-party speech in the clear than either ICSI or Casualtalk, and much more overlap than ICSI, and significantly more than in Casualtalk (Chapter 9 Section 9.2)

Overlap distribution in CasualTalk (Chapter 6 Section 6.2.4) (and TEAMS1) follows a pattern previously observed in ICSI meetings (Laskowski et al., 2012) and casual conversation (Schegloff, 2000), where two-party overlap is much more common than overlap by three or more speakers. In CasualTalk, two-party overlap accounted for 7.6% of conversational time while overlap by three or more only accounted for 0.48%, giving a ratio of two-party to 3+ party overlap of approximately 16:1. In ICSI this ratio was 4.93% to 0.25%, or approximately 20:1 (Laskowski et al., 2012), while in TEAMS1, the ratio was 14.55% to 1.96% or 7.42:1. Comparing the overlap figures for CasualTalk and ICSI, it would appear that overlap is considerably more common in the casual talk examined than in the ICSI meetings corpus, while the ratio of two-party overlap to overlap by three or more speakers is lower. These figures could signal higher interactivity in general in casual speech, with more overlap, and more overlap involving more than two speakers. However, a more generalizable picture of contrasts between the two genres would require matching datasets, with the same speakers engaged in meetings and casual talk, annotated by the same annotation team following the same protocols. In TEAMS1, overlap was higher again than in CasualTalk, suggesting that overlap varies widely with the particular type of conversation under analysis, and not along broad task-based vs. casual lines. However, the number of speakers in CasualTalk is small with many between speaker differences in terms of age, background, and first language. Thus the differences between casual and task based talk will need to be further tested with more data.

Laughter (either alone or overlapping speech) in CasualTalk takes up a total of 8.93% of conversational time (Chapter 6 Section 6.8). The most common situations for laughter were (i) one speaker and one laugher producing (42.76% of the total

conversational time containing laughter), followed by (ii) one participant laughing alone while everyone else is silent (30.36%). Single-party laughter (78.23%) was much more common than shared laughter (21.77%). For shared laughter, 2-party laughter, whether accompanying speech or not, occurs more than three times as often as 3+ party laughter.

Although laughter was present in a similar proportion of conversational time (8.93%) as overlap (8.08%), the ratio of single party laughter to multiparty laughter (3.6:1) was much lower than that of single party speech to overlapping speech (8.5:1). Two-party laughter was more common than 3+ party laughter, with a ratio of 3.4:1, but again the ratio was much lower than that of two-party overlapping speech to 3+ party overlapping speech. Thus laugher is more likely to be shared than speech is, reflecting findings in the literature (Laskowski et al., 2012). Speech and laughter tend to be distributed almost equally among male and female speakers (Chapter 6 Section 6.2.6) in the corpus as a whole.

10.2.2 Composition of Chat and Chunk Phases

This section discusses results of exploration of the characteristics of chat and chunk phases in terms of timing, duration, and distribution, and how chat and chunk phases can be differentiated along simple phase level measures of speech, silence, laughter and overlap in conversation. The results discussed below are summarized in Table 10.1

All conversations contained more chunk than chat phases and more time spent overall in chunk than in chat (Chapter 6 Section 6.3.1). All participants contributed chunks, with the percentage of chunks contributed by a single participant ranging from 58.5% for a participant in conversation B to 11.1% by a participant in Conversation A (Chapter 6 Section 6.3.2).

In terms of the distribution of chat and chunk phases across conversations, more

Feature	Result	Significance
Duration	chunks longer	$p < .01$
Silence	more likely in chunks	$p < .001$
	length of silences greater in chunks	$p < .001$
Overlap	more likely in chat	$p < .001$
Laughter	more likely in chat	$p < .001$

Table 10.1 – Results on differentiation of chat and chunk phases in terms of duration, incidence of silence, overlap and laughter, showing p-values for significant differences found.

chat was seen at conversation beginnings, with chat predominating for the first 8-10 minutes of conversations (Chapter 6 Section 6.3.3). Although our sample size in terms of number of conversations 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. The likelihood of a chunk transitioning to chat is higher in the first part of conversations, but then the likelihood of chunk transitioning to chunk grows and after 10 minutes there is a stronger likelihood of sequential chunks occurring, the familiar 'story-swopping' of long casual conversations. There was more chunk speech in CasualTalk than reported by Slade for her workplace conversation data, 64% in CasualTalk compared to 50% in Slade's data, probably due to the longer length of the CasualTalk conversations, c. 1 hour each compared to 15 minutes in Slade's data, reflecting the higher frequency of chunks as conversations continued.

It is interesting to speculate why casual conversations develop in this way, with less interactivity as the conversation continues. Possibly, the maintenance of a conversational game board with multiple contributing participants would be mentally very costly for long periods; perhaps, once social grooming of the type described by Dunbar has been accomplished by the initial interactive chat, the longer chunk phases allow for participants to take responsibility for larger stretches of time, entertaining or informing their interlocutors. This patterning, and particularly the high

incidence of chunk phases also demonstrates further that conversation is not usually a series of questions and answers, and, particularly during chunks, reinforces the idea that the floor is not always viewed as a scarce resource in casual talk, and that turn competition is not the norm.

For both Slade and Eggins' work and the work described here, the data came from very diverse speakers, and thus individual speaker characteristics may have an effect on the outcomes of experiments. However, the chat and chunk results and insights from the Slade and Eggins data and the CasualTalk dataset are quite similar. Further investigation with more data is needed to ensure that effects noticed are not due to individual speakers.

Durations for chunks (geometric mean: 34 seconds) were significantly higher than for chat segments (28 seconds), and chunk durations showed a strong central tendency with mean (geometric) of 34 seconds (Chapter 6 Section 6.4). There was no significant difference found in chunk durations due to conversation, gender, or speaker, although significant differences were found between duration distributions for chats from different conversations. This elasticity in the length of chat could reflect a possible function as the 'glue' or 'matrix' between chunks in longer conversations, where chat continues until a chunk is established. The stronger central tendency in chunk duration results suggest that there is a preferred length of time for one speaker to dominate in a casual conversation. No correlation was found between the duration of chat or chunk phases and their distance from beginning of conversation, indicating that chunks in particular, and by extension, narratives may have a preferred or standard duration in conversation.

Single party speech occupied the greatest proportion of conversation time in both chat and chunk phases, followed by silence. Global silences were found to be both more frequent and longer in duration in chunk than in chat phases (Chapter 6 Section 6.5). Overlapped speech takes up a far greater proportion of floor time in chat

(17.03%) than in chunk (7.13%) phases (Chapter 6 Section 6.5.3). A Wilcoxon Rank Sum Test with continuity correction shows significant difference in the proportion of overlapping speech to total speech in chat vs chunk phases ($p<.001$), with significantly greater proportions of overlap in chat than in chunk phases. Mean overlap in chat phases was 16% (median: 14.3%), while in chunk mean overlap was 7% (median:5.7%). Overlap involving more than 2 speakers is very rare in both chat and chunk phases. Interestingly, the rarity of more than two speakers talking concurrently was noted in recent work on turn distribution in multiparty storytelling (Rüuhlemann and Gries, 2015) – the results on CasualTalk would seem to show the same phenomenon in casual conversation, where it much more likely for a speaker to be overlapped by one other speaker than by two or more others.

Laughter was found to be significantly more common in chat than in chunk phases (Chapter 6 Section 6.5), with almost twice as much laughter produced by participants in chat (9.5% of all vocal production by participants) as in chunk (4.9%). Who was laughing was interesting in chunk phases, where most laughter was produced by non-chunk owners (81.6%), over four times as much as that produced by chunk owners in chunks (18.4%), reflecting the prevalence of 'audience' reaction to the chunk owner's production.

The differences found between chat and chunk phases in terms of overlap and laughter, in addition to those in duration and frequency of global silences, and those found in the proportions of silence and single party speech in the comparison with the TEAMS corpus discussed below in Section 10.4 indicate that chat and chunk phases can be clearly differentiated for the data studied using simple measures of speech, silence, laughter and overlap. The creation of a chronemic model of casual talk using speech and silence timings and trained on annotated chat and chunk is a very interesting avenue for further work.

10.2.3 Disfluency

This section discusses results on the distribution of disfluencies in casual conversation and in chat and chunk phases.

After applying the Shriberg's Type Classification Algorithm, CasualTalk Conversation A was found to contain 210 instances of filled pauses in or adjacent to speech (plus 185 isolated filled pauses), 280 of repetitions, 211 of deletions, 70 of substitutions, 9 insertions, and 5 hybrid cases (Chapter 7 Section 7.2.1). Shriberg reports that the highest rates of disfluency in her work were seen for filled pauses (FP), repetitions (REP) and deletions (DEL), which made up 85% of the disfluency tokens found in the Switchboard corpus. Disfluency in CasualTalk follows a similar pattern, with 89% of disfluencies accounted for by filled pauses (excluding isolated filled pauses), repetitions, and deletions.

Disfluencies by speakers other than the chunk owner during chunks total 39, by chunk owners during chunks total 434, while disfluencies by all speakers during chat total 289 (Chapter 7 Section 7.3). In chunks, chunk owners hold the floor with minimal input from other speakers, while in chat all participants can participate equally. In order to contrast these two modalities, the 39 disfluencies by non-chunk owners during chunks were excluded.

There was not a great difference in the distributions, but deletions were proportionally more common in chat phases while filled pauses were proportionally more common in chunk phases overall. In view of the very small sample of speakers, the distributions for each speaker were examined to see if the differences observed overall were present for all speakers. Although the proportions varied for individual speakers, in all cases, filled pauses were also proportionately higher in chunk owner speech versus other speech, and repetitions were also proportionally higher in chunk owner speech for each speaker.

The mean duration of filled pauses in chunk owner speech was longer than any of

the other conditions (Chapter 7 Section 7.3.1). The mean durations of filled pauses in the vicinity of utterances by the same speaker is very similar for non-chunk owners in chunk and for speakers in chat. These f-type pauses for everyone but chunk owners are shorter than isolated pauses produced in chat or by non-chunk owners in chunks. However, these results are based on small numbers of speakers and whether isolated pauses are shorter in each speaker varies.

The reduction in the frequency of deletions in chunk vs chat in the conversation is interesting. A reduction in deletions in this modality probably reflects the lack of completion for turns – there could be fewer false starts than would occur when more than one speaker is trying to secure the floor for a turn, as can be seen for all speakers in chat, where there are more deletions. The higher number of filled pauses in chunk owner speech is interesting - they could reflect 'thinking' pauses, when the chunk owner is 'thinking aloud'.

The results for disfluency do not show very strong differentiation between chat and chunk based on disfluency alone (Chapter 7 Section 7.3.2), largely due to the small numbers involved, but do show some tendencies which make further investigation interesting, if suitable data and annotation were available.

10.3 Dynamics

10.3.1 Chat and Chunk

This section discusses results relevant to Research Question 3, **Can chat and chunk phases be differentiated at the utterance or local to turntaking level in terms of duration and type of utterance, silence, overlap and laughter?**. The results discussed below are summarized in Table 10.2

The length of utterances by participants in chat and chunk were contrasted after splitting participant utterances into three classes, Chunk Owner in chunk, Non-Chunk Owner in chunk, and Chat Speaker (Chapter 8 Section 8.2). Chunk Owner

Feature	Result	Significance
Utterance length	longest in chunk owner utterances	p < .001
Laughter length	longer in chat	p < .001
Overlap	two-party overlap intervals longer in chat	p < .001
Liveliness	compression higher in chat	p < .001

ß

Table 10.2 – Results on differentiation of chat and chunk phases in terms of utterance level features and dynamics, showing p-values for significant differences found.

utterances were significantly longer than both Chat Speaker utterances (p<2.2e-16) and Non-Chunk Owner utterances (p<2.2e-16), while Chat Speaker utterances were also significantly longer than Non-Chunk Owner utterances (p<1.2e-16).

For laughter the situation was reversed for Chunk Owners, with Chunk Owner laugher vocalisations significantly shorter than Chat Speaker laughs (p=0.003), and also significantly shorter than Non-Chunk owner laughs (p=0.005). There was no significant difference in laugh length between Chat Speakers and Non-Chunk owners.

The duration of two-party intervals of overlap were compared in chat and chunk phases, as two-party overlap is by far the most common overlap type in the data. Two party overlap duration was contrasted for chat and chunk phases using a one-tailed Wilcoxon Rank Sum test, and two-party overlap duration in chat phases was found to be significantly higher than in chunk phases (W = 2332700, p-value = 0.0001053), at mean 420 ms in chat and 360 in chunk. This difference may reflect the higher likelihood of competition for the floor in chat, with several speakers attempting to contribute at the same time, while the shorter duration in chunk may point to the greater frequency of backchannelling. Wilcoxon Rank Sum Test on global silence length in chat and chunk phases showed that silences in chunk phases were significantly longer than in chat phases (W = 18601000, p-value = 1.562e-05).

The analysis showed that more solo speaker utterances ended in silence than overlap in both chat and chunk phases, although chat sequences, as expected, show more overlap and between-speaker activity. Within speaker silences were very com-

mon in the data, and some such silences may be a function of breathing. High quality annotated recordings would be needed to further explore this question. .

10.3.2 Liveliness

This section discusses results on the characteristics of casual talk and its constituent chat and chunk phases in terms of 'liveliness' or speech dynamics.

Two ways of measuring liveliness were explored. – the existing compression measure and the novel floor change rate measures introduced in this thesis (Chapter 8 Section 8.4). All of the measures showed greater liveliness in chat than in chunk segments when measured at the conversation level. The measures based on speech only for compression and floor change rate were calculated for all individual chat and chunk phases and measures for chat were shown to be significantly higher than those for chunk phases.

The overall compression measures for the conversation in terms of speech (1.1) was slightly higher than that found by Burger and Laskowski for three different sets of recordings of multiparty interaction - interactive seminar segments from the CHIL corpus (1.06), coffee break segments from the CHIL corpus (1.03) and the ICSI meetings corpus (1.08). The chat speech compression rate in CasualTalk (1.18) was higher again, while the chunk ratio of 1.08 was also as high or higher than all three of the corpora measured by Burger and Laskowski, perhaps indicating that casual talk, and particularly chat is highly interactive, with a lot of overlapping speech.

The overall compression in laughter in CasualTalk (1.28) was lower than that found for the interactive seminar segments from the CHIL corpus (1.46), coffee break segments from the CHIL corpus (1.5) and the ICSI meetings corpus (1.7). The rates for chat (1.33) and chunk (1.23) were also lower. This initially surprising result could be due to differences in how speech laughs were treated in the Laskowski and Burger study, or, in the case of the ICSI corpus due to the larger number of participants in

meetings. The distributions of speech compression rates for chat and chunk segments were compared and speech compression in chat was found to be significantly higher than that for chunk (W = 55373, p-value < 2.2e-16).

Floor state change rates were measured for the data set as a whole at an average 1.3 per second for speech (1.4 for speech and laughter). Laughter was treated as silence for floor remaining change rate calculations. Overall rates for chat were 1.53 per second and 1.22 per second for chunk for speech. The distributions of floor change rates per second for each phase were compared, and the rates for chat were found to be significantly higher than for chunk (W = 51475, p-value = 1.927e-12). The TEAM1 dataset had a floor change rate of 1.27 per second, showing it to be considerably less lively than chat, and slightly more lively than chunk phases.

In summary, compression and floor change rates were significantly higher in chat than in chunk, reflecting higher interactivity in chat phases. The compression result indicates more overlap while the floor state change rate result shows more frequent contributions from more speakers, thus providing two complementary views of the liveliness or level of interactivity in talk.

10.3.3 Within and between speaker transitions

This section discusses results on the characteristics of within and between speaker transitions around silence and overlap in casual conversations and how chat and chunk phases differ in terms of within and between speaker transitions.

Bigram transitions reflected the composition of the corpus, with the majority of transitions involving the entry or exit of one speaker from the floor state (Chapter 8 Section 8.5.2). Thus, the most common transitions occurred between one-party speech and silence, followed by transitions between m-party speech and m+1-party overlap, with decreasing frequency with increasing m, again reflecting the finding that overlap is overwhelmingly 2-party, and the rarity of 3+ person overlap in the data.

The results on transition n-grams between intervals of one speaker speaking in the clear for at least one second (***1Sp1***) show that chat and chunk differ in that between speaker transitions in chat interaction are spread over more intervening intervals than in chunk, thus increasing the frequency of more complex transitions (Chapter 8 Section 8.6). This could reflect more turn competition, or indeed more backchannels and acknowledgement tokens being contributed by more participants. Within speaker transitions are predominantly one-interval, perhaps reflecting breathing pauses. One-interval transitions comprise the largest class, with a higher proportion of one-interval transitions in chunk than chat, and higher proportions of within speaker than between speaker one-interval transitions in both, but particularly in monologic chunk. A very interesting finding was that less than 50% of transitions happened with only one intervening interval of silence or overlap, and that most transitions therefore were more complex than the within and between speaker silences and overlaps used in previous work on two-party conversation.

The comparison in labelling of ***1Sp1***-***1Sp1*** and ***1Sp1***-***1SpAny*** transitions starting with the same ***1Sp1*** shows that a significant proportion of labels change depending on how transitions are defined (Chapter 8 Section 8.6.3). This, and the finding that more than half of all transitions involve more than one intervening interval, point to the need for clear and standard definitions and specifications for interspeaker activity.

10.4 Comparison with TEAMS Corpus

This section discusses results on how casual talk and chat and chunk phases compare to task-based conversation in terms of speech, silence, overlap, liveliness, and conditions around floor state changes. The chat and chunk phases in the CasualTalk dataset were compared with TEAMS1, a collection of collaborative games from the TEAMS Corpus (Chapter 9 Section 9.1.1).

The comparison showed significant differences between the proportional distri-

bution of silence, single-party speech and overlap in the three types of multiparty spoken interaction - game, chat and chunk (Chapter 9 Section 9.2). Game interactions had significantly more silence (median 27.%) than Chat (median 24.6%), while Chat segments in turn had significantly more silence than Chunk segments (median 20.9%).

There was significantly more 1-party speech in Chunk (median 72.3%) than in Chat (median 63.2%), and in Chat than in Game conditions (median 53.9%). Overlap was significantly higher in Game (median 14.5%) than in Chat (median 11%), and in Chat than in Chunk (4.7%).

In terms of the two liveliness measures used in the thesis the TEAMS1 dataset had an average floor change rate of 1.27 per second, with chunks averaging 1.22 floor changes per second and chat showing an average floor state change rate of 1.5 per second. The average compression (speech only) for TEAMS1 was 1.26, compared to 1.084 for chunks, and 1.163 for chat (Chapter 9 Section 9.3).

There were differences and similarities in the distribution of how single-speaker stretches end in the three conditions, giving insight into transitions from one floor state to another (Chapter 9 Section 9.4). As with the findings for chat and chunk in Chapter 8, game interaction followed the same general distribution of transitions, with one-interval transitions as the most common class, at 35% of all game transitions.

In general, the game and chat conditions are most similar in transition terms, although they can be distinguished by the significantly higher silence quotient in the game. This may be due to time spent in thought during games while chat is more consistently interactive. There was also a noticeably higher percentage of one-interval between speaker transitions in the game condition than in either chat or chunk, perhaps reflecting quick turn changes as players engaged in motivated discussions. The lower level of single-party speech in the clear and higher level of overlap in Teams

could also indicate intense discussion, consistent with the need to arrive at a solution through verbal interaction

Chunk segments of casual talk contained less silence and overlap and more single party speech than either of the other conditions, reflecting the dominance of one speaker.

Once again, the choice of duration threshold for the right hand single party speech interval in transitions strongly affected how transitions were categorized (Chapter 9 Section 9.4.3), with the overall level of confusion in labelling in ***1Sp1-1Sp1*** transitions sharing their left-hand ***1Sp1*** intervals with ***1Sp1-1SpAny*** transitions in game (27.51%) very close to that seen in chat (27.87%), although the confusion in game transitions was higher in Between Speaker transitions (16.84%) and lower in Within Speaker transitions (10.67%) than that found in chat (13.56% Between and 14.31% Within).

10.5 Insights for artificial dialogue technology

Greater understanding of the distribution of speech, silence, and overlap in chat and chunk phases could aid in the design of dialog systems and active listening systems. Systems which could handle both types of interaction would require adaptive endpointing and turntaking mechanisms tuned to both chat and chunk modalities, in order to convincingly handle the two phases. For example, a system would need to produce feedback when a user was in chunk mode, interact more fully in chat mode, and produce prosodically realistic chunks, with utterances and pauses following the parameters learned from human-human data.

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.

The significant differences in the frequency and duration of silences between chat and chunk phases indicate the need for adaptable endpointing mechanisms to calculate whether and when a casual talk system should start speaking after detecting a silence. These endpointing systems would help generate human like prosody, especially important in non-task based talk. The longer silences in chunk phases pose a particular problem, as a 'one size fits all' endpointing module using a silence threshold suitable for chat could be seen as interruptive during chunk phases. The distribution and duration results for chat and chunk phases will impact the design of dialog systems. 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 there is a need to model the chunk heavy central segments of conversation in order to create systems which form a longer-term dialogic relationship with users. The results on chunk duration, and the 34 second mean duration for chunks could prove useful in the design of dialog systems, guiding the system's 'dosing' of information to conform with the dimensions of information rich single party chunks. This could prove particularly useful for educational or informative applications, or applications which require narrative or storytelling skills.

The differences shown between the three speech exchange systems contrasted here (CasualTalk Chat, CasualTalk Chunk, and TEAMS1 games), demonstrate the importance of understanding the differences in structure and dynamics of types of interaction, as artificial systems trained on game data will not necessarily be accurate in predicting the dynamics of conversational speech. Such knowledge has two main applications. Firstly, it signals the need for data collection in areas of spoken conversation beyond short chats or tasks. Secondly, by identifying similarities in aspects of different interaction types, cataloguing of the dynamics of different interaction types would in help in choosing relevant data from a variety of sources to partially model speech exchange systems where sufficient data is not available for the exact domain.

10.6 Conclusion

This chapter has summarised and discussed the results of the experiments in previous chapters. In the next chapter, conclusions are drawn, limitations and impact are discussed, and future work is outlined.

Chapter 11

Conclusions

The work described in this thesis comprised an examination of the current understanding of multiparty casual conversation, and several corpus studies performed in order to form a more accurate picture of its overall characteristics, chat and chunk structure, and conversation dynamics. In this chapter, conclusions are drawn from the results of the experimental work, impact and limitations are discussed, and future avenues of research are outlined.

The examination of the general structure of the casual conversations in CasualTalk supported accounts in the literature of such talk's structure of chat and chunk phases. The annotation process showed that chat and chunk phases can be annotated with a high degree of inter-annotator agreement, by human annotators following high level definitions in the literature. Experiments have shown that chat and chunk phases differ significantly in terms of low level features - distribution of speech and silence, proportion of laughter and overlap, distribution of transition types, and to a lesser extent in disfluency. There are also clear differences in the liveliness of the two sub-genres. These results make the creation of a stochastic chronemic classifier based on speech and silence a clear avenue of future work. Such a model could automatically distinguish between chat and chunk phases in conversation. In comparison

with the task-based TEAMS data, there were significant differences in the distribution of silence and overlap, pointing to the need to consider and analyse different genres or speech exchange systems separately to form an accurate picture of their dynamics.

The different phases of casual talk will need to be considered in the design and implementation of artificial dialogue systems which can provide a realistic casual or social conversation experience for users. The timing information, greater understanding of the constitution of casual talk's phases, and knowledge of the differences in activity at the end of utterances by a single speaker point to the need for systems which can adapt to different phases or speech exchange systems, in order to provide human-like interaction beyond simple question answering.

The work on within and between speaker transitions was particularly interesting, as the findings that more than half of all transitions are complex point to the need for more analysis of what exactly is happening around turn change and retention. There is clear scope for future work involving further classification of transitions depending on the number of distinct speakers involved, and investigation of the duration of transitions, as well as for similar studies of other corpora, and a larger number and wider variety of speakers.

The major challenge in the work described was the scarcity of suitable data and the need for manual segmentation and annotation, which was extremely time consuming. This, combined with the small number of speakers examined could weaken the results found, as they could be influenced by the characteristics of individual speakers. The small number of conversations was also a limitation, although the bulk of the work focussed on chat and chunk segments which were present in numbers sufficient for robust statistical analysis, as confirmed by results showing no effect of conversation and speaker on distributions. The CasualTalk dataset in particular contains very diverse speakers, reinforcing the possibility of speaker effects on results. An additional limitation is that only English language corpora were studied, and thus

effects due to a particular language could not be investigated. The literature on differences in turntaking practices between languages has produced mixed results. For example, broad similarities have been found in question-answer turn changes in over 10 languages with local differences in gap and overlap duration (Stivers et al., 2009), while some ethnographic studies have found differences in the distribution of overlap in turn changes (Martinez, 2018). Recent work using the methods introduced in this thesis on Swedish and Estonian casual conversation data reported chat and chunk patterns (Aare et al., 2020), and found floor change dynamics similar to those found in the English data studied in this thesis (Gilmartin et al., 2020), providing some evidence that these phenomena may be widespread. The application of the methods used here to more languages, speaker populations and situations is a very promising avenue for future work.

The results found could be useful in the design of more human-like dialog systems for a range of applications, particularly those requiring users to converse with the system in entertainment, education, or healthcare settings.

In conclusion, this thesis has described quantitative investigations of multiparty casual talk, a speech genre which is fundamental to human social life. Modelling such conversations will aid in building convincing artificial dialogue. Further analysis in this area will prove fruitful but depends on the availability of suitable datasets. Understanding the dynamics of task based interaction has been greatly aided by efforts to record high quality corpora, and it is hoped that these explorations will strengthen the case for the production of high quality multiparty casual conversation corpora.

Appendix A

Words Added to CMU Dictionary for Alignment

- AFFORDANCES AH0 F AO1 R D EH2 N S AH0 Z
- ANONYMIZE AH0 N AA1 N AH0 M AY2 Z
- ASSOCIATIONISM AH0 S OW2 S IY0 EY1 SH AH0 N IH2 Z AH0 M
- BILINGUALITY B AY0 L IH1 NG G W AH0 L IH2 T IY0
- BILINGUALLY B AY0 L IH1 NG G W AH0 L IY0
- BIOMETRICALLY B AY1 OW0 M EH1 T R IH0 K L IY0
- CATHA'S K AA0 TH ER0 Z
- COILING K OY1 L IH0 NG
- CONEMARA K AA2 N AH0 M AA1 R AH0
- DODGY D AA1 JH IY0
- DUBBERS D AH1 B ER0 Z

- DUZAN D UW1 Z AH0 N
- DVD D IY2 V IY2 D IY1
- DVDS D IY2 V IY2 D IY1 Z
- ENSCHEDE EH1 N SH EH0 D EY2
- ERS ER0 Z
- EUROS Y UW1 R OW2 Z
- FLEECES F L IY1 S AH0 Z
- GUTTURALS G AH1 T ER0 AH0 L Z
- HAARLEM HH AA1 R L AH0 M
- HEMA HH EY1 M AH0
- HILVERSUM H IH1 L V ER0 S UH2 M
- HOODIES HH UH0 D IY1 Z
- HOUSEY HH AW1 S IY0
- IONTACH IY1 AH0 N T UH0 K
- KILLINEY K IH2 L AY1 N IY0
- KNACKERS N AE1 K ER0 Z
- LADEFOGED L AE1 D EH0 F OW2 G IH0 D
- LAOGHAIRE L EH1 R IY0
- LARPING L AA1 R P IH0 NG
- LEAT L AE1 TH

- LUAS L UW1 AH0 S
- MAITH M AY1
- MISE M IH1 SH EH0S
- MM AH1 M
- MM-HM AH1 M HH AHI M
- HM HH AH1 M
- NESCAFE N EH1 S K AH0 F EY2
- NOUNS N AW1 N Z
- ODDNESS AA1 D N AH0 S
- OPTI AA1 P T IY0
- ORTHOGRAPHICALLY AO0 R TH AA1 G R AH0 F IH0 K L IY0
- PHONETICIANS F AH0 N EH1 T IH0 SH AH0 N Z
- PHONOLOGY F AH0 N AA1 L AA1 JH IY2
- POITIN P AH1 CH IY2 N
- PREPOSITIONS P R AH0 P AH0 Z IH1 SH AH0 N Z
- REBOOTING R IY0 B UW1 T IH0 NG
- SCHENGEN SH EH1 NG AH0 N
- SEGMENTAL S EH2 G M EH1 N T AH0 L
- SHITE SH AY1 T
- TALLAGHT T AE1 L AH0

- THORNTON'S TH AO1 R N T AH0 N Z
- TOLKIEN T AA1 L K IY0 N
- TWENTE T W EH1 N T EY1
- UNEXPRESSIVE AH2 IH0 K S P R EH1 S IH0 V
- UTRECHT UW0 T R EH1 K T
- VEG V EH1 JH
- WOAH W OW1
- YOHASHI Y OW1 H AE2 SH IY0

Bibliography

K. Aare, E. Gilmartin, M. Wlodarczak, P. Lippus, and M. Heldner. Breath Holds in Chat and Chunk Phases of Multiparty Casual Conversation. In *Proc. 10th International Conference on Speech Prosody 2020*, pages 779–783, 2020. doi: 10.21437/SpeechProsody.2020-159. URL `http://dx.doi.org/10.21437/SpeechProsody.2020-159`.

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

J. Allen, G. Ferguson, and A. Stent. An architecture for more realistic conversational systems. In *International Conference on Intelligent User Interfaces: Proceedings of the 6th International Conference on Intelligent User Interfaces*, volume 14, pages 1–8, New York, NY, 2001a. Association for Computing Machinery.

J. F. Allen, D. K. Byron, M. Dzikovska, G. Ferguson, L. Galescu, and A. Stent. Toward conversational human-computer interaction. *AI Magazine*, 22(4):27, 2001b.

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

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

J. Ang, R. Dhillon, A. Krupski, E. Shriberg, and A. Stolcke. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*, Denver, CO, 2002. International Speech Communication Association (ISCA).

O. Aran, H. Hung, and D. Gatica-Perez. A multimodal corpus for studying dominance in small group conversations. In *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, volume 22, Valletta, Malta, 2010.

A. J. Aubrey, D. Marshall, P. L. Rosin, J. Vandeventer, D. W. Cunningham, and C. Wallraven. Cardiff Conversation Database (CCDb): A Database of Natural Dyadic Conversations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition – Workshops (CVPRW), 2013*, pages 277–282, Portland, OR, June 2013.

Audacity. Audacity: Free Audio Editor and Recorder, 2014.

P. Auer. A conversation analytic approach to code-switching and transfer. In M. Heller, editor, *Codeswitching: Anthropological and sociolinguistic perspectives*, volume 48, pages 187–213. DeGruyter Mouton, Boston, 1988.

J. Austin and J. Urmson. *How to do things with words*. Harvard University Press, 1978.

J. A. Bachorowski and M. J. Owren. Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological Science*, 6(4):219–224, 1995.

R. Baker and V. Hazan. LUCID: a corpus of spontaneous and read clear speech in British English. In *Proceedings of the DiSS-LPSS Joint Workshop 2010*, pages 3–6, Tokyo, Japan, 2010.

R. Baker and V. Hazan. DiapixUK: task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, 43(3):761–770, 2011.

M. Bakhtin. The problem of speech genres. In *Speech genres and other late essays*, pages 60–102. University of Texas Press, 1986.

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

M. E. Beckman and J. Hirschberg. The ToBI annotation conventions. *Ohio State University, Technical Report*, 1994.

B. Beebe, D. Alson, J. Jaffe, S. Feldstein, and C. Crown. Vocal congruence in mother-infant play. *Journal of Psycholinguistic Research*, 17(3):245–259, 1988.

B. Benwell and M. McCreaddie. Keeping small talk small in health-care encounters: Negotiating the boundaries between on- and off-task talk. *Research on Language and Social Interaction*, 49(3):258–271, 2016.

C. R. Berger. Communication theories and other curios. *Communications Monographs*, 58(1):101–113, 1991.

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

T. Bickmore, D. Schulman, and L. Yin. Maintaining engagement in long-term interventions with relational agents. *Applied Artificial Intelligence*, 24(6):648–666, 2010.

D. W. Black. Laughter. *JAMA: the journal of the American Medical Association*, 252(21): 2995–2998, 1984.

BNC Consortium. *British National Corpus*. BNC Consortium, 2000.

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

P. T. Brady. A technique for investigating on-off patterns of speech. *Bell System Technical Journal*, 44(1):1–22, 1965.

P. T. Brady. A Statistical Analysis of On-Off Patterns in 16 Conversations. *Bell System Technical Journal*, 47(1):73–91, 1968.

P. T. Brady. A Model for Generating On-Off Speech Patterns in Two-Way Conversation. *Bell System Technical Journal*, 48(7):2445–2472, 1969.

S. E. Brennan and E. A. Hulteen. Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems*, 8(2-3):143–151, 1995.

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

M. Bull and M. Aylett. An analysis of the timing of turn-taking in a corpus of goal-oriented dialogue. In *Proceedings of the 5th International Converence on Spoken Language Processing: Australian Speech Science and Technology Association*, Canberra, Australia, 1998.

S. Burger, K. Laskowski, and M. Wolfel. A Comparative Cross-Domain Study of the Occurrence of Laughter in Meeting and Seminar Corpora. In *Proceedings of The Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco, 2008. European Language Resources Association.

F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss. A database of german emotional speech. In *Ninth European Conference on Speech Communication and Technology*. ISCA, International Speech Communication Association (ISCA), 2005.

R. Burling. *The talking ape: How language evolved*. Oxford University Press, USA, 2007.

N. Campbell. Conversational speech synthesis and the need for some laughter. *IEEE Transactions on Audio, Speech, and Language Processing,*, 14(4):1171–1178, 2006.

N. Campbell. On the use of nonverbal speech sounds in human communication. In *Proceedings of the 2007 COST action 2102 international Conference on Verbal and Nonverbal Communication Behaviours*, pages 117–128, Dublin, Ireland, 2007.

N. Campbell. An audio-visual approach to measuring discourse synchrony in multimodal conversation data. *Linguistic Theory and Raw Sound*, 40:199, 2009.

A. Canavan, D. Graff, and G. Zipperlen. *Callhome American English Speech Corpus*. Linguistic Data Consortium (LDC), Philadelphia, PA, 1997.

L. Cassotta, S. Feldstein, and J. Jaffe. The stability and modifiability of individual vocal characteristics in stress and nonstress interviews. *William Alanson White Institute Research Bulletin*, 2, 1967.

O. Cetin and E. Shriberg. Speaker overlaps and asr errors in meetings: Effects before, during, and after the overlap. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2006*, volume 1, pages I–I, Toulouse, France, 2006. IEEE.

W. Chafe. *Discourse, consciousness, and time: The flow and displacement of conscious experience in speaking and writing*. University of Chicago Press, 1994.

W. Chafe and J. Danielewicz. Properties of spoken and written language. In R. Horowitz and S. Samuels, editors, *Comprehending oral and written language*, pages 83–113. Academic Press, San Diego, CA, 1987.

C. Cheepen. *The predictability of informal conversation*. Pinter Publishing, London, UK, 1988.

N. Chomsky. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA, 1965.

B. Christian. *The Most Human Human*. Doubleday, New York, 2011.

H. Clark. *Using language*. Cambridge University Press, 1996.

J. Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

J. Coleman, M. Liberman, G. Kochanski, L. Burnard, and J. Yuan. Mining a year of speech. *VLSP 2011: New Tools and Methods for Very-Large-Scale Phonetics Research*, pages 16–19, 2011.

E. Couper-Kuhlen and M. Selting. *Prosody in conversation: Interactional studies*. Cambridge University Press, 2006.

M. Dalton and A. Ní Chasaide. Modelling intonation in three irish dialects. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS*, volume 1, pages 1073–1076, Barcelona, Spain, 2003.

M. E. Dawson, A. M. Schell, and D. L. Filion. The Electrodermal System. In J. T. Cacioppo, L. G. Tassinary, and G. G. Berntson, editors, *Handbook of psychophysiology*, pages 159–181. Cambridge University Press, 2007.

J. Deese. Pauses, prosody, and the demands of production in language. In *Temporal Variables in Speech*. Mouton Publishers, Boston, MA, 1980.

J. W. Du Bois, W. L. Chafe, C. Meyer, S. A. Thompson, and N. Martey. *Santa Barbara Corpus of Spoken American English*. Linguistic Data Consortium (LDC), Philadelphia, PA, 2000.

R. Dunbar. The Social Brain: Mind, Language, and Society in Evolutionary Perspective. *Annual Review of Anthropology*, 32(1):163–181, 2003.

R. I. M. Dunbar. *Grooming, gossip, and the evolution of language*. Harvard University Press, 1998a.

R. I. M. Dunbar. Theory of mind and the evolution of language. In J. R. Hurford, M. Studdert-Kennedy, and C. Knight, editors, *Approaches to the Evolution of Language*, pages 92–110. Cambridge University Press, 1998b.

S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283, 1972.

J. Edlund. *In search for the conversational homunculus: serving to understand spoken human face-to-face interaction.* PhD thesis, KTH Royal Institute of Technology, 2011.

J. Edlund, J. Beskow, K. Elenius, K. Hellmer, S. Strömbergsson, and D. House. Spontal: A Swedish Spontaneous Dialogue Corpus of Audio, Video and Motion Capture. In *Proceedings of The Language Resources and Evaluation Conference (LREC 2010)*, Valletta, Malta, 2010. European Language Resources Association.

S. Eggins and D. Slade. *Analysing casual conversation.* Equinox Publishing, London, 2004.

R. Eklund. *Disfluency in Swedish human-human and human-machine travel booking dialogues.* PhD thesis, Linköping University, Linköping, Sweden, 2004.

M. Ernestus and R. H. Baayen. Corpora and exemplars in phonology. In J. A. Goldsmith, J. Riggle, and C. Alan, editors, *The Handbook of Phonlogical Theory*, pages 374–400. Wiley-Blackwell, Oxford, 2011.

M. Ernestus, L. Kočková-Amortová, and P. Pollak. The nijmegen corpus of casual czech. In *LREC 2014: 9th International Conference on Language Resources and Evaluation*, pages 365–370, 2014.

M. T. C. Ernestus. *Voice assimilation and segment reduction in casual Dutch, a corpus-based study of the phonology-phonetics interface.* Utrecht: LOT, 2000.

F. Eyben, F. Weninger, F. Gross, and B. Schuller. Recent Developments in openSMILE, the Munich Open-source Multimedia Feature Extractor. In *Proceedings of the 21st ACM International Conference on Multimedia*, MM '13, pages 835–838, New York, NY, USA, 2013a. ACM. 00218.

F. Eyben, F. Weninger, S. Squartini, and B. Schuller. Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013*, pages 483–487, Vancouver, Canada, 2013b. IEEE.

N. Fairclough. *Discourse and social change*. Polity Press, Cambridge, 1992.

S. Feldstein and J. Welkowitz. A chronography of conversation: In defense of an objective approach. In W. Siegman and S. Feldstein, editors, *Nonverbal behavior and communication*, pages 435–499. Erlbaum, Hillsdale, NJ, 1978.

R. Fernández and J. Ginzburg. Non-sentential utterances in dialogue: A corpus-based study. In *Proceedings of the 3rd SIGdial workshop on Discourse and dialogue-Volume 2*, pages 15–26, Philadelphia, USA, 2002. Association for Computational Linguistics.

I. R. Finlayson. *Testing the roles of disfluency and rate of speech in the coordination of conversation*. PhD thesis, Queen Margaret University, 2014.

J. G. Fiscus, J. Ajot, M. Michel, and J. S. Garofolo. The rich transcription 2006 spring meeting recognition evaluation. In *International Workshop on Machine Learning for Multimodal Interaction*, pages 309–322, Bethesda, ML, 2006. Springer.

E. Gilmartin, M. O'Reilly, C. Saam, B. R. Cowan, C. Vogel, N. Campbell, and V. Wade. Silence and overlap in chat and chunk phases of multiparty casual conversation. In *Proc. 9th International Conference on Speech Prosody 2018*, pages 379–383, 2018.

E. Gilmartin, K. Aare, M. O'Reilly, and M. Wlodarczak. Between and Within Speaker Transitions in Multiparty Conversation. In *Proc. 10th International Conference on Speech Prosody 2020*, pages 799–803, 2020. doi: 10.21437/SpeechProsody.2020-163. URL http://dx.doi.org/10.21437/SpeechProsody.2020-163.

L. H. Gilpin, D. M. Olson, and T. Alrashed. Perception of speaker personality traits using speech signals. In *CHI EA '18: Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–6, Montreal, Canada, 2018. ACM.

J. Ginzburg. *The interactive stance.* Oxford University Press, 2012.

J. Ginzburg, Y. Tian, P. Amsili, C. Beyssade, B. Hemforth, Y. Mathieu, C. Saillard, J. Hough, S. Kousidis, and D. Schlangen. The Disfluency, Exclamation and Laughter in Dialogue (DUEL) Project. In *Proceedings of the 18th SemDial Workshop on the Semantics and Pragmatics of Dialogue (DialWatt), Edinburgh*, 2014.

J. Gleick. *The information: A history, a theory, a flood.* Harper Collins, UK, 2011.

P. J. Glenn. *Laughter in interaction.* Cambridge University Press, 2003.

J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 1992*, volume 1, pages 517–520, San Francisco, CA, 1992.

E. Goffman. *Interaction ritual.* Anchor Books, New York, 1967.

F. Goldman-Eisler. *Psycholinguistics: Experiments in spontaneous speech.* Academic Press, San Diego, CA, 1968.

E. Grabe. The IViE labelling guide. *Journal of the Acoustical Society of America*, 101: 3728–3740, 2001.

A. C. Graesser, P. Chipman, B. C. Haynes, and A. Olney. AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, 48 (4):612–618, 2005.

B. Granström. Towards a virtual language tutor. In *Proceedings of the InSTIL/ICALL*

Symposium on NLP and Speech Technologies in Advanced Language Learning 2004, pages 1–8, Venice, Italy, 2004.

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

H. P. Grice. Logic and conversation. In P. Cole and J. L. Morgan, editors, *Syntax and Semantics 3: Speech acts*, pages 26–40. Academic Press, San Diego, CA, 1975.

J. J. Gumperz. Contextualization revisited. In P. Auer and A. Di Luzio, editors, *The contextualization of language*, volume 22, pages 39–53. John Benjamins Publishing Amsterdam, 1992.

J. Hagedorn, J. Hailpern, and K. G. Karahalios. VCode and VData: illustrating a new framework for supporting the video annotation workflow. In *Proceedings of the working conference on Advanced Visual Interfaces*, pages 317–321, Palermo, Italy, 2008.

M. A. K. Halliday. *Spoken and written language*. Oxford University Press, USA, 1989.

M. A. K. Halliday. Spoken and written modes of meaning. *Media texts: Authors and readers*, 7:51–73, 1994.

S. I. Hayakawa. *Language in thought and action*. Harcourt Brace Jovanovich, San Diego, 1990.

M. Heldner and J. Edlund. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568, Oct. 2010.

M. Heldner, J. Edlund, K. Laskowski, and A. Pelcé. Prosodic features in the vicinity of silences and overlaps. In *Proceedings of the 10th Nordic Conference on Prosody*, pages 95–105, Helsinki, Finland, 2008.

S. Hennig, R. Chellali, and N. Campbell. The D-ANS corpus: the Dublin-Autonomous Nervous System corpus of biosignal and multimodal recordings of conversational speech. In *Proceedings of The Language Resources and Evaluation Conference (LREC 2014)*, pages 3438–3443, Reykjavik, Iceland, 2014. European Language Resources Association.

J. Holmes. Doing collegiality and keeping control at work: Small talk in government departments. In *Small talk*, pages 52–81. Routledge, UK, 2014.

E. Holt. The last laugh: Shared laughter and topic termination. *Journal of Pragmatics*, 42(6):1513–1525, 2010.

E. Holt. Laughter at Last: Playfulness and laughter in interaction. *Journal of Pragmatics*, 100:89–102, 2016.

R. Ishii and Y. I. Nakano. Estimating user's conversational engagement based on gaze behaviors. In *Proceedings of the International Workshop on Intelligent Virtual Agents 2008*, pages 200–207, Tokyo, Japan, 2008. Springer.

A. V. Ivanov, G. Riccardi, A. J. Sporka, and J. Franc. Recognition of personality traits from human spoken conversations. In *INTERSPEECH 2011*, pages 1549–1552, Florence, Italy, 2011. ISCA, International Speech Communication Association (ISCA).

J. Jaffe, L. Cassotta, and S. Feldstein. Markovian model of time patterns of speech. *Science*, 144(3620):884–886, 1964.

J. Jaffe, S. Feldstein, and L. Cassotta. Markovian models of dialogic time patterns. *Nature*, 216(5110):93–94, 1967.

J. Jaffe, B. Beebe, S. Feldstein, C. L. Crown, M. D. Jasnow, P. Rochat, and D. N. Stern. Rhythms of dialogue in infancy: Coordinated timing in development. *Monographs of the society for research in child development*, 66(2):i–149, 2001.

R. Jakobson. Closing statement: Linguistics and poetics. In T. A. Sebeok, editor, *Style in language*, pages 350–377. MIT Press, Cambridge, MA, 1960.

A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, and A. Stolcke. The ICSI meeting corpus. In *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing, ICASSP 2003*, volume 1, pages 364–368, Hong Kong, China, 2003.

G. Jefferson. A technique for inviting laughter and its subsequent acceptance/declination. In G. Psathas, editor, *Everyday Language: Studies in Ethnomethodology*, chapter 4, page 79–96. Irvington, New York, NY, 1979.

S. Johnson. A Dictionary of the English Language, 1755.

W. Johnson. Measurements of oral reading and speaking rate and disfluency of adult male and female stutterers and nonstutterers. *The Journal of speech and Hearing disorders. Monograph Supplement*, 7:1–20, 1961.

K. Jokinen. Gaze and gesture activity in communication. In *Universal Access in Human-Computer Interaction. Intelligent and Ubiquitous Interaction Environments*, pages 537–546. Springer, 2009.

K. Jokinen. Turn taking, utterance density, and gaze patterns as cues to conversational activity. In *Proceedings of ICMI 2011, Workshop Multimodal Corpora for Machine Learning: Taking Stock and Roadmapping the Future*, pages 31–36, Alicante, Spain, 2011.

J. Kelly and J. Local. *Doing phonology: observing, recording, interpreting.* Manchester University Press, 1989.

W. T. B. Kelvin. *Popular lectures and addresses*, volume 1. Macmillan and Company, London, 1891.

A. Kendon. *Conducting Interaction: Patterns of behavior in focused encounters*, volume 7. Cambridge University Press, 1990.

M. Kipp. Anvil - A Generic Annotation Tool for Multimodal Dialogue. In *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech)*, pages 1367–1370, Aalborg, Denmark, 2001.

J. Krause, C. Lalueza-Fox, L. Orlando, W. Enard, R. E. Green, H. A. Burbano, J.-J. Hublin, C. Hönni, J. Fortea, and M. De La Rasilla. The Derived FOXP2 Variant of Modern Humans Was Shared with Neandertals. *Current Biology*, 17(21):1908–1912, 2007.

I. Kruijff-Korbayová, E. Oleari, I. Baroni, B. Kiefer, M. C. Zelati, C. Pozzi, and A. Sanna. Effects of off-activity talk in human-robot interaction with diabetic children. In *Proceedings of ROMAN 2014, the 23rd IEEE International Symposium on Robot and Human Interactive Communication*, pages 649–654, Edinburgh, Scotland, 2014. IEEE.

J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

K. Laskowski. *Predicting, detecting and explaining the occurrence of vocal activity in multi-party conversation.* PhD thesis, Carnegie Mellon University, 2011.

K. Laskowski. On the conversant-specificity of stochastic turn-taking models. In *INTERSPEECH 2014*, Interspeech, pages 2026–2030, Singapore, 2014. ISCA, International Speech Communication Association (ISCA).

K. Laskowski. A framework for the automatic inference of stochastic turn-taking styles. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 202–211, Los Angeles, CA, 2016.

K. Laskowski and S. Burger. Analysis of the occurrence of laughter in meetings. In *INTERSPEECH 2007*, Interspeech, pages 1258–1261, Antwerp, Belgium, 2007. ISCA, International Speech Communication Association (ISCA).

K. Laskowski and Q. Jin. Modeling Prosody for Speaker Recognition: Why Estimating Pitch May Be a Red Herring. In *Proceedings of Odyssey 2010*, pages 12–19, Brno, Czech Republic, 2010.

K. Laskowski and T. Schultz. Modeling vocal interaction for segmentation in meeting recognition. In *Machine Learning for Multimodal Interaction*, pages 259–270. Springer, 2007.

K. Laskowski and T. Schultz. Recovering participant identities in meetings from a probabilistic description of vocal interaction. In *INTERSPEECH 2008*, Interspeech, pages 82–85, Brisbane, Australia, 2008. ISCA, International Speech Communication Association (ISCA).

K. Laskowski and E. Shriberg. Modeling other talkers for improved dialog act recognition in meetings. In *INTERSPEECH 2009*, Interspeech, pages 2783–2786, Brighton, UK, 2009. ISCA, International Speech Communication Association (ISCA).

K. Laskowski, M. Heldner, and J. Edlund. On the dynamics of overlap in multi-party conversation. In *INTERSPEECH 2012*, Interspeech, pages 846–849, Portland, OR, 2012. ISCA, International Speech Communication Association (ISCA).

J. Laver. Communicative functions of phatic communion. In *Organization of behavior in face-to-face interaction*, pages 215–238. De Gruyter Mouton, Boston, MA, 1975.

J. L. Lemke. Analyzing verbal data: Principles, methods, and problems. In *Second International Handbook of Science Education*, pages 1471–1484. Springer, 2012.

W. J. Levelt. Monitoring and self-repair in speech. *Cognition*, 14(1):41–104, 1983.

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

H. G. Liddell, R. Scott, H. S. Jones, and R. McKenzie. *A Greek English Lexicon*. Clarendon Press, Oxford, UK, 1966.

D. Litman, S. Paletz, Z. Rahimi, S. Allegretti, and C. Rice. The TEAMS corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431, Austin, Texas, 2016.

D. J. Litman and S. Silliman. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Demonstration Papers at HLT-NAACL 2004*, pages 5–8, Boston, MA, 2004.

D. Little. The Common European Framework of Reference for Languages: Content, purpose, origin, reception and impact. *Language Teaching*, 39(03):167–190, 2006.

Y. Liu, C. Fox, M. Hasan, and T. Hain. The Sheffield Wargame Corpus — Day Two and Day Three. In *INTERSPEECH 2016*, pages 3833–3837, San Francisco, CA, 2016. International Speech Communication Association (ISCA).

J. Local and G. Walker. Mind the gap': further resources in the production of multi-unit, multi-action turns. *York Papers in Linguistics*, 3:133–143, 2005.

J. Lyons. *Semantics. vol. 2*. Cambridge University Press, 1977.

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

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

C. B. Martinez. Cross-cultural analysis of turn-taking practices in english and spanish

conversations. *Vernacular: New Connections in Language, Literature, & Culture*, 3 (1):5, 2018.

J. D. Matarazzo, M. Weitman, G. Saslow, and A. N. Wiens. Interviewer influence on durations of interviewee speech. *Journal of Verbal Learning and Verbal Behavior*, 1 (6):451–458, 1963.

J. D. Matarazzo, A. N. Wiens, G. Saslow, B. V. Allen, and M. Weitman. Interviewer Mm-Hmm and interviewee speech durations. *Psychotherapy: Theory, Research & Practice*, 1(3):109, 1964.

N. Mattar and I. Wachsmuth. Small talk is more than chit-chat. In *KI 2012: Advances in Artificial Intelligence*, pages 119–130. Springer, Berlin, Germany, 2012.

D. W. Maynard and P. L. Hudak. Small talk, high stakes: Interactional disattentiveness in the context of prosocial doctor-patient interaction. *Language in Society*, 37(5): 661–688, 2008.

M. McCarthy. Mutually captive audiences: Small talk and the genre of close-contact service encounters. In *Small talk*, pages 84–109. Routledge, UK, 2014.

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

M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3): 276–282, 2012.

M. Mehu and R. I. M. Dunbar. Relationship between smiling and laughter in humans (Homo sapiens): Testing the power asymmetry hypothesis. *Folia Primatologica*, 79 (5):269–280, 2008.

M. Mehu, K. Grammer, and R. I. M. Dunbar. Smiles when sharing. *Evolution and Human Behavior*, 28(6):415–422, 2007.

G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63(2):81, 1956.

P. Müller, M. X. Huang, and A. Bulling. Detecting low rapport during natural interactions in small groups from non-verbal behaviour. In *23rd International Conference on Intelligent User Interfaces*, pages 153–164, Tokyo, Japan, 2018. ACM.

J. Nichols. The origin and dispersal of languages: Linguistic evidence. *The origin and diversification of language*, (24):127–170, 1998.

A. C. Norwine and O. J. Murphy. Characteristic time intervals in telephonic conversation. *Bell System Technical Journal*, 17(2):281–291, 1938.

OED. Oxford english dictionary, 2020. URL `https://www.oed.com/view/Entry/40748?rskey=LPln79&result=1`. "conversation, n.".

C. Oertel and G. Salvi. A gaze-based method for relating group involvement to individual engagement in multimodal multiparty dialogue. In *Proceedings of the 15th International Conference on Multimodal Interaction*, pages 99–106, Sydney, Australia, 2013. ACM.

C. Oertel, F. Cummins, J. Edlund, P. Wagner, and N. Campbell. D64: A corpus of richly recorded conversational interaction. *Journal on Multimodal User Interfaces*, 7:19–28, 2013.

R. Ogden. The phonetics of talk in interaction–introduction to the special issue. *Language and Speech*, 55(1):3–11, 2012.

D. R. Olson. From utterance to text: The bias of language in speech and writing. *Harvard Educational Review*, 47(3):257–281, 1977.

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

P. Paggio, J. Allwood, E. Ahlsén, K. Jokinen, and C. Navarretta. The NOMCO Multimodal Nordic Resource-Goals and Characteristics. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC 2010)*, pages 19–21, Valletta, Malta, 2010. European Language Resources Association.

C. Perreault and S. Mathew. Dating the origin of language using phonemic diversity. *PloS one*, 7(4):e35289, 2012.

M.-Z. Poh, N. C. Swenson, and R. W. Picard. A wearable sensor for unobtrusive, long-term assessment of electrodermal activity. *IEEE Transactions on Biomedical Engineering*, 57(5):1243–1252, 2010.

R. R. Provine. Laughing, tickling, and the evolution of speech and self. *Current Directions in Psychological Science*, 13(6):215–218, 2004.

R. R. Provine, R. J. Spencer, and D. L. Mandell. Emotional expression online: Emoticons punctuate website text messages. *Journal of Language and Social Psychology*, 26(3):299–307, 2007.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

A. Raux, D. Bohus, B. Langner, A. W. Black, and M. Eskenazi. Doing research on a deployed spoken dialogue system: one year of Let's Go! experience. In *INTERSPEECH 2006*, Interspeech, Pittsburgh, PA, 2006. ISCA, International Speech Communication Association (ISCA).

P. Reichl and F. Hammer. Hot discussion or frosty dialogue? Towards a temperature metric for conversational interactivity. In *Proceedings of the Eighth International Conference on Spoken Language Processing*, Jeju, South Korea, 2004.

D. Reitter, J. D. Moore, and F. Keller. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, pages 685–690, Vancouver, Canada, 2006.

T. Rietveld, M. Ernestus, et al. Pitfalls in corpus research. *Computers and the Humanities*, 38(4):343–362, 2004.

C. Rüuhlemann and S. Gries. Turn order and turn distribution in multi-party storytelling. *Journal of Pragmatics*, 87:171–191, 2015.

H. Sacks, E. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.

A. Sapru and H. Bourlard. Detecting speaker roles and topic changes in multiparty conversations using latent topic models. In *INTERSPEECH 2014*, Interspeech, pages 2882–2886, Singapore, 2014. ISCA, International Speech Communication Association (ISCA).

F. D. Saussure. *Course in General Linguistics*. McGraw Hill, New York, 1916.

E. Schegloff. *Sequence organization in interaction: A primer in conversation analysis I*, volume 1. Cambridge University Press, 2007.

E. Schegloff and H. Sacks. Opening up closings. *Semiotica*, 8(4):289–327, 1973.

E. A. Schegloff. Overlapping talk and the organization of turn-taking for conversation. *Language in society*, 29(01):1–63, 2000.

D. Schlangen. What we can learn from dialogue systems that don't work: On dialogue systems as cognitive models. In *Proceedings of DiaHolmia, the 13th International*

Workshop on the Semantics and Pragmatics of Dialogue (SEMDIAL 2009), pages 5–17, Stockholm, Sweden, 2009.

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

J. Searle. *Expression and meaning: Studies in the theory of speech acts*. Cambridge University Press, 1985.

G. Senft. Phatic communion. *Culture and language use*, 2:226–233, 2009.

R. M. Seyfarth, D. L. Cheney, and P. Marler. Vervet monkey alarm calls: semantic communication in a free-ranging primate. *Animal Behaviour*, 28(4):1070–1094, 1980.

C. E. Shannon and W. Weaver. *A mathematical theory of communication*. American Telephone and Telegraph Company, 1948.

Y. Shi, N. Ruiz, R. Taib, E. Choi, and F. Chen. Galvanic skin response (GSR) as an index of cognitive load. In *CHI'07 extended abstracts on Human factors in computing systems*, pages 2651–2656, San Jose, CA, 2007.

E. Shriberg. To 'errrr' is human: ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, 31(1):153–169, 2001. doi: 10.1017/S0025100301001128.

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

G. Skantze. *Higgins Annotation Tool*. KTH, Stockholm, 2009.

D. Slade. *The texture of casual conversation: A multidimensional interpretation*. Equinox, London, 2007.

T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. De Ruiter, K.-E. Yoon, et al. Universals and cultural variation in

turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106 (26):10587–10592, 2009.

D. Tannen. *Spoken and written language: Exploring orality and literacy*. Ablex, Norwood, NJ, 1982.

D. Tannen. Relative focus on involvement in oral and written discourse. *Literacy, language, and learning: The nature and consequences of reading and writing*, pages 124–147, 1985.

The CMU Pronouncing Dictionary. *Version 0.7b*. Carnegie Mellon University, Pittsburgh, PA, 2007.

S. Thornbury and D. Slade. *Conversation: From description to pedagogy*. Cambridge University Press, 2006.

F. Torreira and M. Ernestus. Weakening of intervocalic/s/in the nijmegen corpus of casual spanish. *Phonetica*, 69(3):124–148, 2012.

F. Torreira, M. Adda-Decker, and M. Ernestus. The nijmegen corpus of casual french. *Speech Communication*, 52(3):201–212, 2010.

J. E. F. Tree. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language*, 34(6):709–738, 1995.

J. Trouvain. Segmenting phonetic units in laughter. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 2793–2796, Barcelona, Spain, 2003.

B. V. Tucker and M. Ernestus. Why we need to investigate casual speech to truly understand language production, processing and the mental lexicon. *The mental lexicon*, 11(3):375–400, 2016.

W. Turnbull. *Language in action: Psychological models of conversation*. Routledge, London, UK, 2003.

F. Valente, S. Kim, and P. Motlicek. Annotation and recognition of personality traits in spoken conversations from the AMI meetings corpus. In *INTERSPEECH 2014*, Singapore, 2012. ISCA, International Speech Communication Association (ISCA).

K. J. Van Engen, M. Baese-Berk, R. E. Baker, A. Choi, M. Kim, and A. R. Bradlow. The Wildcat Corpus of native-and foreign-accented English: Communicative efficiency across conversational dyads with varying language alignment profiles. *Language and Speech*, 53(4):510–540, 2010.

L. Vardoulakis, L. Ring, B. Barry, C. Sidner, and T. Bickmore. Designing relational agents as long term social companions for older adults. In N. Y., N. M., P. A., and W. M., editors, *Intelligent Virtual Agents. IVA 2012*, pages 289–302. Springer, Berlin, Germany, 2012.

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

A. Vinciarelli, H. Salamin, A. Polychroniou, G. Mohammadi, and A. Origlia. From nonverbal cues to perception: personality and social attractiveness. In *Cognitive Behavioural Systems*, pages 60–72. Springer, 2012.

R. S. Wallace. The anatomy of A.L.I.C.E. In E. R., R. G., and B. G, editors, *Parsing the Turing Test*, pages 181–210. Springer, Dordrecht, Netherlands, 2009.

M. Warren. *Features of naturalness in conversation*, volume 152. John Benjamins Publishing, Amsterdam, 2006.

P. Watcyn-Jones. *Pair Work 2*. Penguin Books, London, 1981.

E. Waugh. *Brideshead revisited*. Chapman and Hall, London, 1945.

J. Williams, A. Raux, D. Ramachandran, and A. Black. The dialog state tracking challenge. In *Proceedings of SIGDIAL 2013*, pages 404–413, Metz, France, 2013.

J. Wilson. *On the boundaries of conversation*, volume 10. Pergamon, Oxford, UK, 1989.

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

S. J. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. H. Woodland. The hidden Markov modelling toolkit book, version 3.4. *Cambridge University Engineering Department*, 2006.

Z. Yu, A. Papangelis, and A. Rudnicky. TickTock: A Non-Goal-Oriented Multimodal Dialog System with Engagement Awareness. In *2015 AAAI Spring Symposium Series*, Palo Alto, CA, 2015.

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