Toward Summarization of Communicative Activities in Spoken Conversation

John Niekrasz

Doctor of Philosophy
University of Edinburgh
2012

Abstract

This thesis is an inquiry into the nature and structure of face-to-face conversation, with a special focus on group meetings in the workplace. I argue that conversations are composed of episodes, each of which corresponds to an identifiable communicative activity such as giving instructions or telling a story. These activities are important because they are part of participants' commonsense understanding of what happens in a conversation. They appear in natural summaries of conversations such as meeting minutes, and participants talk about them within the conversation itself. Episodic communicative activities therefore represent an essential component of practical, commonsense descriptions of conversations.

The thesis objective is to provide a deeper understanding of how such activities may be recognized and differentiated from one another, and to develop a computational method for doing so automatically. The experiments are thus intended as initial steps toward future applications that will require analysis of such activities, such as an automatic minute-taker for workplace meetings, a browser for broadcast news archives, or an automatic decision mapper for planning interactions.

My main theoretical contribution is to propose a novel analytical framework called participant relational analysis. The proposal argues that communicative activities are principally indicated through participant-relational features, i.e., expressions of relationships between participants and the dialogue. Participant-relational features, such as subjective language, verbal reference to the participants, and the distribution of speech activity amongst the participants, are therefore argued to be a principal means for analyzing the nature and structure of communicative activities.

I then apply the proposed framework to two computational problems: automatic discourse segmentation and automatic discourse segment labeling. The first set of experiments test whether participant-relational features can serve as a basis for automatically segmenting conversations into discourse segments, e.g., activity episodes. Results show that they are effective across different levels of segmentation and different corpora, and

indeed sometimes more effective than the commonly-used method of using semantic links between content words, i.e., lexical cohesion. They also show that feature performance is highly dependent on segment type, suggesting that human-annotated "topic segments" are in fact a multi-dimensional, heterogeneous collection of topic and activity-oriented units.

Analysis of commonly used evaluation measures, performed in conjunction with the segmentation experiments, reveals that they fail to penalize substantially defective results due to inherent biases in the measures. I therefore preface the experiments with a comprehensive analysis of these biases and a proposal for a novel evaluation measure. A re-evaluation of state-of-the-art segmentation algorithms using the novel measure produces substantially different results from previous studies. This raises serious questions about the effectiveness of some state-of-the-art algorithms and helps to identify the most appropriate ones to employ in the subsequent experiments.

I also preface the experiments with an investigation of participant reference, an important type of participant-relational feature. I propose an annotation scheme with novel distinctions for vagueness, discourse function, and addressing-based referent inclusion, each of which are assessed for inter-coder reliability. The produced dataset includes annotations of 11,000 occasions of person-referring.

The second set of experiments concern the use of participant-relational features to automatically identify labels for discourse segments. In contrast to assigning semantic topic labels, such as topical headlines, the proposed algorithm automatically labels segments according to activity type, e.g., presentation, discussion, and evaluation. The method is unsupervised and does not learn from annotated ground truth labels. Rather, it induces the labels through correlations between discourse segment boundaries and the occurrence of bracketing meta-discourse, i.e., occasions when the participants talk explicitly about what has just occurred or what is about to occur. Results show that bracketing meta-discourse is an effective basis for identifying some labels automatically, but that its use is limited if global correlations to segment features are not employed.

This thesis addresses important pre-requisites to the automatic summarization of conversation. What I provide is a novel activity-oriented perspective on how summarization should be approached, and a novel participant-relational approach to conversational analysis. The experimental results show that analysis of participant-relational features is a promising avenue for tackling the difficult problem of activity-oriented summarization.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(John Niekrasz)

Contents

List of Figures

List of Tables

List of Corpus Extracts

Chapter 1

Introduction

Face-to-face conversations are an important part of daily work. At the university institute where this thesis was composed, we have many varieties: student-advisor meetings, seminars, faculty meetings, lectures, study groups, and thesis examinations. Workplace conversations like these are useful because they help people make decisions, share information, and solve problems.

In our institute, what happens in conversation is usually not publicly documented or archived. In faculty meetings, a secretary might take meeting minutes. Video recordings of lectures are occasionally made. Most commonly, however, documentation is left to individuals taking private notes. This lack of public documentation causes information loss. The only recourse for missing a lecture or study group is to rely on someone else's notes (assuming they are understandable and available). If you want to review the highlights from a guest lecture that you missed, you're out of luck.

Individuals in many organizations—the office of a large corporation, for example—typically view such documentation as a hindrance to productivity. Taking meeting minutes is often required in large organizations. However, it is expensive to have a dedicated secretary, so participants are charged with taking notes themselves. This interferes with participation in the meeting and adds work.

The aim of this thesis is to address these problems of conversation documentation by contributing to the development of *systems that produce meeting minutes automatically*. Such a system would automatically produce short descriptions of decisions, actions, and other important events that occurred in a meeting, e.g., "David Thomas presented his analysis of the Q2 budget." It would associate important factual details with each event, e.g., "the Q2 budget will be in deficit by $3m," and each item would be linked to the

Figure 1.1: A hypothetical commitment summary.

Commitment Summary	
assigned to	**Harry**
commitment	**Send Sarah email about travel**
agreed at	**10:34, Wed., 5 June 2007**
by	**Harry, Sarah**
[click here to view recording]	

relevant location in an archived recording of the meeting. This information would then be made available to absentees or attendees wishing to be reminded of what happened.

Consider one type of conversational event that is important to group work and commonly found in meeting minutes—the *commitment to future action*. Business processes rely on them, and if they are remembered incorrectly or forgotten, this causes delay and financial loss. Organizations also need to be able to review commitments for business process evaluation and legal audit. The following is an example of a conversational commitment, taken from a project planning meeting (Burger et al., 2002).

(1) [ISL–m063–28m3s]
01 Sarah: Yeah. also, 'cause you said you were gonna send me an email about how to set up our travel.
02 Harry: Yeah, I'm gonna send- yeah, I'll send you the email uhm uhm when I go back. send you the email. uhm and you're gonna have to contact him, and they have a travel agency.
03 Sarah: Okay.

Having access to a transcript like this could be quite useful. But an even better starting point for anyone needing to review the meeting would be a summary like that shown in Figure 1.1.

As another example, consider the regular meetings of an academic research group. Research groups discuss problems with current experiments, review recent publications, and plan for new grant proposals. Any number of these events might occur in a single meeting, but individuals who miss the meeting might be interested in only the proposal planning. Others might be interested in only the review of publications. The ideal automated minute-taker would thus identify, index, and summarize each of these distinct events separately.

Organizations are eager to make use of meeting minute-takers that might address such requirements. User studies, interviews, and workflow analysis have helped to clarify what users' specific needs are (Lisowska et al., 2004; Whittaker et al., 2006; Pallotta et al., 2007). The aim of this thesis is to develop systems that address these needs in ways that are more effective than current approaches. While it will most likely take many years before computers can perform the analysis required to mimic human minute-taking, this thesis represents important steps toward that objective. In the remainder of this introductory chapter, I shall outline the steps I intend to take and introduce the sub-problems that will be addressed along the way.

1.1 Meeting Summarization as Analysis of Activity Episodes

Meeting minutes may be considered a type of spoken conversation summary, and the area of technological research that is most relevant to the minute-taking task is automatic speech summarization. State-of-the-art methods in automatic summarization typically use the *extractive* paradigm, with a focus on the *subject matter* of the conversation (examples applied to meetings include Zechner (2002); Maskey and Hirschberg (2006); Murray (2008); Xie and Liu (2010); Riedhammer et al. (2010)). The objective of such systems is to identify the most meaningful and important words, phrases, and utterances in the speech transcript, and to present them to the user. This can help users to quickly understand what the conversation is *about*, which in turn helps them perform information access tasks more efficiently (Murray et al., 2009).

One of the problems with this approach to summarization, however, is that the results that are produced can be very difficult to read due to a lack of grammaticality, continuity, and contextualization. Summaries containing only a list of important topical words do not give readers an intuitive understanding of what occurred. Additionally, summaries containing extracts of sentences are often hard to understand without actually listening to a fair amount of surrounding dialogue. Consider Figure 1.2, which shows example output from a state-of-the-art summarizer described in Murray (2008). The employed approach chooses utterances from the meeting whose content words (as opposed to function words) are a minimally redundant set of the words most indicative of the conversation's subject matter. In this case, the example lists the five top-ranking utterances, which have been selected because they contain relevant content words and phrases like "LCD screen," "remote control," and "buttons."

Upon listening to the entire meeting from which this snippet was taken, it is clear

Figure 1.2: An extractive meeting summary (Murray (2008), p. 76).

D: And on top of that the LCD screen would um help in making the remote control easier to use.

B: We've got um the buttons we have to use. The on-off, sound on-off, sound higher or lower, um the numbers, uh zero to uh uh nine. Um the general buttons m more general b one button for shifting up and shifting down uh channel.

D: But if we would make um a changing channels and changing volume button on both sides, that would certainly yield great options for the design of the remote.

A: Uh requirements are uh teletext, docking station, audio signal, small screen, with some extras that uh button information.

D: So they would prefer uh a design where the remote control just lies flat in the docking station.

that this summary adequately 'gists' the main *topics* or *subject matter* of the dialogue. But without having listened to the meeting, there are problems one faces in interpreting the result. It is hard to appreciate what exactly was *happening* when the participants uttered these sentences. The speakers' intentions and motivations are hard to pin down without more context. There are some ways to address such problems. For example, one can link the summary to the original recordings so users can listen back to the necessary context.

This method of speech summarization, however, is also associated with an even more fundamental problem: it does not do *what humans naturally do* when summarizing conversations, nor does it address what people want or need from summaries (Whittaker et al., 2008). People naturally summarize conversations by *abstraction*—they interpret, reduce, and reformulate the conversation. And subject matter is not the only thing they focus on. They also focus on the *activities*, *accomplishments*, and *attitudes* of the participants. This fundamental gap between current methods and natural summaries is a core problem addressed in this thesis. I propose that designing a natural summarization method requires more than an incremental improvement of current methods. It requires a study of natural summaries, natural conversation, and a reconsideration of what users (especially participants themselves) find important in summaries (and what they do with them).

It is therefore of central importance that one consider more than just the technological or computational issues. The sociological, psychological, and anthropological accounts must also be considered, particularly research grounded in the ethnographic tradition, where face-to-face conversation plays a privileged role in the study of social processes. Everyday conversation is in fact fundamental to some influential approaches, including the field of conversation analysis (see Sacks (1992) for an introduction), which argues that conversation is the primary setting in which we create personal and social identities. Levinson (1983, p. 284) also claims that conversation is the "central and most basic" use of language, and that many of the concerns of the field of pragmatics "can be shown to be centrally organized around usage in conversation." Clark (1996) argues that studying conversation is important for several reasons: it is how we learn to use language; it is how we evolved to have language; it is universal to the world's societies. He describes face-to-face conversation as "the one setting that should take priority" (p. 11) if we are ever to universally characterize language use. Employing insights from this research, which deals with the deeper social meanings of language use, is indispensible to achieving the goals of this thesis.

One particularly insightful study is Goffman's (1974) analysis of the organization of human experience. Goffman's goal was to gain an understanding of our subjective experience of everyday interaction, and to begin answering the question that we all answer when encountering a situation—"What is it that's going on here?" He wanted to find out how the structure of an interaction relates to the way we interpret its meaning and purpose. The ideas that he presented—for example, that participants communicate cues that signal their interactional roles—have come to be seen as fundamental to our understanding of interactions.

In some ways, being able to answer Goffman's question is precisely the sort of abstraction that an automatic meeting minute-taker must do, though it is a question that has been largely ignored by current approaches to summarization. All that needs to be done is to transform his question about human experience into a technological question. In other words, given a computer with access to a recorded conversation: how can the computer tell *what it is that's going on*? In particular, we must identify *the participants' subjective experience* of what's going on in the conversation. This question is the most fundamental question addressed in this thesis, and it expresses the main point of view from which the problem of conversational summarization shall be addressed.

There are many directions from which to begin teasing Goffman's question apart, and this study must narrow in on specific issues. In this thesis, the main idea is to give priority to the root purposes for interaction, taking an approach principally informed by the notion that conversation is a *social activity*. That conversation is *social* highlights the fact that it is collaborative (Clark, 1996), socially structured (Hasan, 1991), and organized around roles and relationships between the participants (Hanks, 2005). That conversation is an *activity* puts attention on the idea that it is a goal-directed effort (Grosz and Sidner, 1986) with outcomes and achievements (Clark, 1996). However explicit or implicit, pre-determined or spontaneous, there is an underlying *purpose* for conversation that involves *social action*.

This thesis also argues that if an analysis is to have utility in conversational technologies, then what must be of principal interest are those *activities* that are part of participants' *commonsense*, *subjective* account of a dialogue. For example, a meeting might comprise activities such as giving instructions, disagreeing, committing to a future action, and giving a presentation. These are natural language descriptions of the socially-constituted *activity types* (Levinson, 1992) in which conversation arises, instantiated in summaries to describe specific instances of communicative activity. I shall show later that references to such activities are indeed a principal component of natural summaries, and that their descriptions have a readily discernible structure.

But where should one start with an analysis of such activities in meetings? One obvious point of departure is to think about how they may be identified through observation of the linguistic and interactional structure of a conversation. For example, one might ask when does one activity begin and another end? Consider the following extract from a meeting of an electronics design team working on a new remote control device (Carletta, 2007).

(2) [AMI–ES2008C–878]
855 B: So like one through five, or
856 C: Yeah, yeah
857 D: Like a radio type sorta situation?
858 C: about
859 C: yeah like yeah, a bit like radio presets. Um
860 A: Pre-set channels
861 A: and then we're gonna need um numbers one through zero,
862 A: right?
863 C: Uh we wouldn't even need the numbers.
864 B: No.
865 C: I think maybe numbers seems is kind of old-fashioned.
866 A: Well,
867 A: but in order to pre-set a cha

(2) [cont. . .]

868 A: oh

869 A: I guess you can just hold it down when you get to one when you're scrolling through.

870 C: Yeah, yeah,

871 C: you can just

872 B: Mm.

873 C: and you need some kind of, I dunno, sort of up down kind of button,

874 B: Yeah, up down.

875 C: but the volume control could double for that, for example.

876 A: Mm-hmm.

877 A: Okay, um

878 A: finishing the meeting now.

879 A: Um our next meeting starts in thirty minutes,

880 A: um you each have things to do, look and feel design, user interface design, product evaluation,

881 A: and you two are going to work together on a prototype using modelling clay.

882 A: You'll get specific instructions from your personal coach.

883 B: Ooh.

884 C: Cool.

885 D: Wow.

886 A: Um did we decide on a chip?

887 A: Let's go with a simple chip?

888 B: Simple chip.

889 C: Yep.

890 A: Okay.

891 A: We are done.

892 A: Thank you everyone.

In the first section of this dialogue until line 876, several individuals are discussing the pros and cons of remote control design ideas. Following line 876, there is an abrupt shift into what might be called a 'wrap-up' in which the team leader is reviewing final decisions, responsibilities, and future actions. These two episodes are the manifestation of two separate activities. In a meeting minutes document, each activity should be labeled with its own summary description.

This thesis proposes that many of the conversational activities a minute-taker should be interested in will commonly occur, like above, as cohesive *episodes* within a conversation (Korolija, 1998b). Korolija notes that episodes naturally emerge from participants' efforts to maintain coherent social interaction, and this produces a readily discernible intermediate level of dialogue structure—greater than a single speech act but shorter than an entire conversation. The dialogue above is a good example of two distinct episodes. On either side of line 876, there is an element of coherence and similarity amongst the utterances. At the boundary, there are identifiable markers of a transition.

What we are searching for then are spans of a conversation, each associated with a discernible activity. To discover the locations of the episodes and the properties of the activities is this work's principal analytical task, and the central challenge is to understand how these are indicated by linguistic and interactional features.

1.2 Starting Points and Conjectures

To begin drawing connections between activities and their linguistic indicators, it is helpful to first draw a parallel between this problem and that addressed by Grosz and Sidner's theory of discourse structure (Grosz and Sidner, 1986). In their theory, discourse is structured principally around the intentions (or purposes) of the participants. Discourse segments—the units of the linguistic structure—are manifestations of the intentional structure, and they arise by virtue of the participants addressing a particular purpose for a period of time. This thesis argues that the relationship between activities and episodes is similar to this, and a parallel can be drawn between the elements of an episodic activity analysis and those in the Grosz and Sidner theory. In the remainder of this thesis, episodes are thus often referred to as a type of *discourse segment*. Activities are considered part of the underlying intentional structure of a conversation and are modeled around a dominant purpose.

Looking again at the example dialogue **(2)**, one can search for indications of a purpose in each episode. The first episode contains an evaluation "I think maybe. . . " [865] and a suggestion ". . . we're gonna need. . . " [861]. The use of evaluations and suggestions by multiple participants indicates that some sort of collaborative planning activity is occurring. One might call this example a 'brainstorming' activity. This can be contrasted with, for example, a 'personal narrative' activity, which is unlikely to contain this configuration of utterance types. The second episode contains references to the future and confirmations of responsibilities. These are commonly associated with activities like 'decision-making' and meeting 'wrap-ups' (contrasting with, for example, a 'presentation' or 'report').

The occurrence of different intentionally-typed expressions in each episode helps to distinguish each from the other, and helps to identify some of the properties of the underlying activities. It is even possible to glean an impression of the activity while neglecting the order of utterances within such an episode. The distribution (in the statistical sense) of utterance types provides enough of an indication of what's going on. This kind of shallow-structured patterning, where consistent distributions of expression types are found within episodes, is the basis for my first main conjecture.

1.2.1 Conjecture 1: Participant-relational coherence for segmentation

Participant-relational coherence is a proposed concept based upon Halliday and Hasan's (1976) notion of cohesion. Cohesion refers to the relations between elements of a text that give it coherence and meaning. Lexical cohesion, a variety of such relations that concerns semantic links between words, has proved one of the most effective means for automatic discourse segmentation of conversation (Galley et al., 2003; Eisenstein and Barzilay, 2008). The method goes back at least to Morris and Hirst (1991), whose contribution was to show that lexical cohesion is stronger within a discourse segment than between segments. Identifying lexical relations thus allows identification of the segments. In our example, the words *numbers* and *preset* are good examples of cohesive words that might help to distinguish the first segment from the latter.

But one problem with relying on semantic lexical patterns is that such patterns do not themselves constitute a theoretically-grounded foundation of discourse segment structure. Rather, discourse theory suggests they are a reflection of an underlying structure, and factors such as intentions (Grosz and Sidner, 1986) and joint activity (Clark, 1996) are more fundamental to the origin of observable linguistic patterns. This raises the question of whether segmentation can be done using alternatives to lexical-semantic features that are more directly indicative of activities or intentions themselves. For example, one might look to expressions like those identified in our example above, such as "I think maybe" [865]. The words in this phrase would be considered 'stop words' by a typical lexical-semantic method, and they would be automatically removed from analysis.

This thesis proposes a novel solution—a synthesis of theoretical arguments called *participant-relational analysis*. What participant-relational analysis proposes is that activity types may be distinguished and analyzed in terms of *participant-relational features*, i.e., those features of a discourse that are evidence of *relationships* between participants and the dialogue. Participant-relational analysis is inspired by theoretical work including Chafe's (1994) work on conscious experience in speaking, Wiebe's (1994) work on point-of-view in narratives, and linguistic anthropological work on deixis and reference (Hanks, 1990). Participant-relational features include, for example, attitudinal (i.e., subjectivity, sentiment) and perspectival (i.e., temporal, aspectual, deictic) relationships, the level of speaker activity, and references to participants. In essence, the approach relies on the principle that each activity is the realization of a unique configuration of participants, roles, attitudes, and relationships—a framework or schema for interaction (Goffman, 1974).

The first conjecture I put forth is that the utterances within a given activity episode will exhibit consistency of participant-relational features. For example, in the brainstorming example one observes several trends. Participants use evaluative expressions such as "I think" or "old-fashioned." They use certain modals, semi-modals, and auxiliaries like "gonna", "wouldn't", and "maybe" to express specific aspectual and attitudinal relationships to the content. They make reference to the decision-makers using references such as *we* and *our*. Finally, I propose that these relations are expressed throughout the episode in an identifiably consistent way.

1.2.2 Conjecture 2: Bracketing meta-discourse for segment labeling

Returning to the example dialogue, consider now the transition between episodes. Lines 877 and 878 are the first two lines of the 'wrap-up' episode, and they contain important indicators of the start of a new episode. First, there is the discourse marker "Okay, um" [877]. It has been demonstrated that discourse markers like *well* and *okay* have a relationship to discourse segment boundaries (Hirschberg and Litman, 1993; Passonneau and Litman, 1997; Bangerter et al., 2004) and may be effectively used in automatic segmentation methods (Galley et al., 2003; Eisenstein and Barzilay, 2008; Hsueh, 2008b).

In contrast to line 877, however, the phrase "finishing the meeting now" in line 878 provides explicit information indicating what is about to follow. This is an example of *bracketing meta-discourse* (Schiffrin, 1980), i.e., discourse about discourse that has just occurred or is about to occur. Another (hypothetical) example of such meta-discourse is "why don't you go ahead and present." This kind of language has a privileged place in our problem because it reflects the participants' own conception of what is going on in a conversation. It addresses the problem of getting at participants' subjective experience of the dialogue.

The current thesis places bracketing meta-discourse in the context of participant-relational analysis. The second conjecture I put forth is that bracketing meta-discourse (which may be seen as a unique kind of participant-relational expression) may be used to identify properties of adjacent conversational activities. For example, the participant-relational deictic expression "now" in "finishing the meeting now" [878] helps to signal its relationship to the next activity, and "finishing the meeting" identifies the essential character of that activity. And the participant reference "you" in "why don't you go ahead and present" helps to identify roles in the ensuing presentation activity. The conjecture proposes that the meta-discourse that 'brackets' (Schiffrin, 1980) episodes of activity by occurring near episode boundaries will contain useful explicit information about those activities. The descriptive content of these utterances, e.g., "present" and "finishing the meeting," if it can be identified and interpreted correctly, could be useful as a means for automatic identification of descriptive labels for activities.

1.3 Research Problems, Methods, and Objectives

I have structured this thesis into two main parts, addressing the problem of meeting summarization (i.e., meeting minute-taking) from two distinct angles. In the first part, I provide a broad, explorative justification to the approach just introduced. The objective here is to synthesize existing theories from a wide range of disciplines and provide a convincing argument in support of (1) an activity-oriented approach and (2) a participant-relational analysis. To articulate what I mean by episodic communicative activities and participant relations, I shall refer often to examples and conduct a preliminary quantitative analysis.

In the second part of the thesis, I apply participant-relational analysis to computational experiments on corpora of annotated conversations. The objective of this second part is to advance technological understanding and improve the state-of-the-art in conversation summarization. I do not provide a complete solution to the summarization problem, nor do I create a completely automated minute-taker. Rather, I tackle two necessary pre-requisites: the segmentation of conversations by activity episode, and the identification of activity-oriented labels for those episodes.

1.3.1 Outline of Chapters 2 and 3: Qualitative Research

The exploratory stage of research is covered by Chapters 2 and 3. In Chapter 2, I introduce a framework for analyzing summarization systems proposed in Spärck Jones (1999, 2001, 2007). I then employ this framework to review the literature, with the purpose of identifying trends in research, synthesizing some general conclusions, and identifying avenues for future progress. The following are some of the questions that are addressed in Chapter 2.

What is summarization?
What are the main components of summarization systems?
How are these related to information extraction?
What are the historical trends in speech summarization?
What lessons can be learned from previous research?
Which directions are promising for the future?
How does this thesis relate to previous work?

Chapter 3 is where I lay out support for an activity-oriented approach to conversation summarization. My argument is based upon two main sources: theories of conversational structure from multiple disciplines, and specific analyses of real conversations and summaries. I draw principally from the theoretical work of Goffman (1974); Clark (1996); Grosz and Sidner (1986); Levinson (1983); Hasan (1991) and Korolija (1998b). The empirical analysis draws mainly from the AMI corpus of workplace meetings (Carletta, 2007), though the Pear Stories corpus of conversational monologues (Chafe, 1994) is also studied to support generalization. Chapter 3 addresses the following primary research questions:

To what ends do participants use conversation?
Which types of activities show up in minutes documents?
How are communicative activities interactionally realized?
What properties of activities should be included in a summary?

In the last section of Chapter 3, I describe the participant-relational method of conversation analysis, which is a synthesis of theoretical analyses in discourse analysis (Chafe, 1994; Wiebe, 1994; Schiffrin, 1980), linguistic anthropology (Hanks, 1990), and ethnomethodology (Goffman, 1974). The goal here is to show in more concrete terms how communicative activities are indicated by specific linguistic and interactional features. Participant-relational analysis thus addresses the following primary research questions:

> Which features provide a basis for distinguishing communicative activities?
> How are activity types indicated by verbal contribution?
> What level of complexity is captured by such features?
> How might this be translated into an automated algorithm?

1.3.2 Outline of Chapters 4 and 5: Quantitative Research

The purpose for developing participant-relational analysis is to motivate novel *quantitative, computational* approaches to automatic activity-oriented summarization. In the second part of the thesis, I conduct experiments toward this end. This quantitative stage of research is composed of two main parts, covered by Experiments 1 through 4, and Experiments 5 and 6. The aim here is to apply participant-relational analysis to two essential sub-problems in automatic conversation summarization: episode-level discourse segmentation and activity type labeling.

Before presenting the experiments, however, the thesis presents two important prerequisites having to do with validating the experimental designs. First, we must understand the reliability of any annotations or content coding that must be performed. Second, we must confirm the appropriateness of any evaluation measures used. This thesis devotes Chapter 4 to these two issues. First, Section 4.1 describes a novel method for *annotating references to people* in conversation. Understanding reference to persons is important because little is known about how reliably participant-relational features can be identified in texts, and such features depend heavily upon the resolution of person references. The annotation study is therefore designed principally to test the reliability of coding person reference information in conversations. The study also contributes generally to the understanding of participant deixis and the problem of annotating vague and ambiguous referring expressions. Section 4.2 presents a novel mathematical analysis of commonly-used *segmentation evaluation methods*, showing that such methods are substantially flawed due to biases toward certain types of segmentation. State-of-the-art segmentation algorithms are therefore re-evaluated and a new segmentation evaluation measure is proposed. This study thus contributes to understanding of the effectiveness of state-of-the-art methods, and identifies appropriate algorithms for use in the subsequent experiments. Chapter 4 addresses the following primary research questions:

> What are the obstacles to effective coding of participant-relational features?
> How does ambiguity and vagueness influence participant reference resolution?
> How are segmentations evaluated?
> Are the dominant evaluation methods effective and unbiased?
> What is the state-of-the-art in automatic segmentation?

The first part of Chapter 5 (Experiments 1 through 4) contains a set of experiments on *discourse segmentation*. The experiments test whether participant-relational features can serve as a basis for segmenting conversations, and contribute to knowledge of the most effective methods for identifying episodes in conversations. Experiment 1 addresses the problem of fine-grained intentional segmentation of conversational monologues, expanding a notable previous study (Passonneau and Litman, 1997). Experiment 2 then presents an initial study comparing the participant-relational approach to the traditional topic-based approach. Experiment 3 conducts an initial test of the participant-relational method on multi-party meeting dialogues, comparing its effectiveness with state-of-the-art methods. Finally, Experiment 4 addresses the segmentation problem in its greatest detail through studying the effect that different activity types have on the performance of various segmentation methods. This contributes a deeper, more nuanced understanding of the nature of *both human and automated* discourse segments in meeting conversations. These experiments collectively address the following primary research questions:

> Do participant-relational features indicate fine-grained intentional segments?
> Do participant-relational features indicate coarse-grained topic segments?
> Do participant-relational features indicate episodes of communicative activity?
> Do participant-relational approaches perform as well as the state-of-the-art?
> How do various approaches relate to the segment activity type?

The second part of Chapter 5 (Experiments 5 and 6) presents experiments on automatic identification of activity labels (e.g., presentation, discussion, evaluation) using bracketing meta-discourse. Experiment 5 is designed to extract a set of labels for all activity types in a corpus. Experiment 6 presents an algorithm for assigning these labels to individual activity episodes. These experiments thus address the following primary research questions:

> Can meta-discourse be used to identify a set of activity type labels for a corpus?
> Can meta-discourse be used to label specific activity episodes?

Finally, Chapter 6 presents a summary of the presented results and a discussion of promising avenues for future work.

Chapter 2

Trends in Automatic Speech Summarization

In this chapter, I review the literature on automatic speech summarization. The purpose of the survey is to identify current trends in speech and conversation summarization. I shall pay special attention to techniques that are not focused exclusively on subject matter, and I will focus on factors associated with discourse and interaction. This will provide a historical and methodological context for my activity-oriented approach, and it will provide motivation for the segmentation and labeling experiments presented in subsequent chapters.

It is common in introductions to summarization to begin by explaining the typological distinction between *extractive* and *abstractive* summaries (Spärck Jones, 1996), which specifies whether a summary contains either fragments from the input source (extractive), or a novel reformulation of its content (abstractive). It is also common to refer to the distinction between *indicative* and *informative* summaries (Borko and Bernier, 1975). Indicative summaries *refer* users to the content and help users identify important elements in the source documents. Informative summaries actually *convey* the important contents themselves, and are instead a stand-in for the source.

There is a problem, however, with relying upon basic typological distinctions when studying summarization. Summarizers typically reside in the grey area between the extremes of such categories, and they employ such a diverse array of inputs, processes, outputs, and uses, that a thorough comparison requires a richer framework for system description. There are systems that summarize movies, books, people, events, and images. Summaries can be presented as video recordings, paragraphs of text, or lists of web pages. Summaries can contain topical headlines, indexes into the source, or descriptions of the author's attitudes. And for different genres of source material, a variety of genre-specific summary types can be identified, such as a *synopsis* of events in a narrative, or a *precis* of an author's arguments (Rowley, 1982). And in relation to any one of these factors, summarization systems might employ any number of possible realizations. This diversity makes it difficult to identify common components across summarizers, and especially difficult to pin down exactly which categories can be applied. This means that a rich descriptive framework is a necessary first step toward identifying specific problems with current systems and relating them to this thesis.

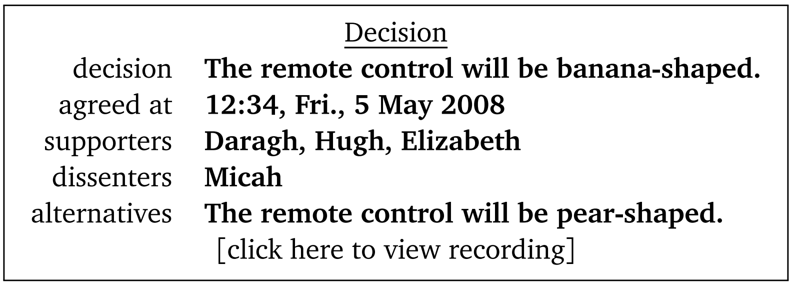
To assist in painting a more nuanced picture, it is helpful to draw upon a framework for characterizing summarization systems developed by Spärck Jones (1999, 2001, 2007). Her framework is especially useful because it focuses on *natural language* summarization systems, and it captures the many changes in research in this area over the last decade. She divides system description into three main areas: *system structure*, *summarizing factors*, and *evaluation*, and she establishes a conceptual and terminological basis for teasing apart subtle distinctions. This framework shall serve as a foundation for my review, in which I will describe and compare state-of-the-art systems, highlight current problems in meeting summarization, and motivate the problems addressed in this thesis.

2.1 Describing and Evaluating Summarization Systems

I begin by defining summarization, using one idea that stands out as a unifying concept—that a summarization system somehow *reduces* or *condenses* its source material. Spärck Jones (1999) uses this as the basis for her definition of **summarization** as a "reductive transformation [. . .] through content condensation [. . .]." Maybury (1995) characterizes it as "distilling the most important information from a source or set of sources to produce an abridged version [. . .]." From these definitions, it is intuitive to conclude that summarization may be also be defined upon a notion of *utility*—as a type of process that aids rapid consumption of information that would otherwise demand more resources to consume.

From this description, it is evident that summarization has a close relationship to **information extraction** (Cowie and Lehnert, 1996; Cowie and Wilks, 2000), in which specific facts are extracted from natural language text. This is particularly true for applications that are specialized for a particular domain. In automatic meeting analysis, for example, there are systems that extract the decisions (Fernández et al., 2008b; Hsueh, 2008b) or action items (Purver et al., 2007) in a meeting. Even though these systems do not attempt to

Figure 2.1: A hypothetical meeting decision summary.



Decision

decision	**The remote control will be banana-shaped.**
agreed at	**12:34, Fri., 5 May 2008**
supporters	**Daragh, Hugh, Elizabeth**
dissenters	**Micah**
alternatives	**The remote control will be pear-shaped.**

[click here to view recording]

reduce *all* the important information in their source, they nonetheless summarize certain types of information in the source. This highlights the fact that users often consider the 'important' parts of a document to be limited to specific types of information, in which case, the distinction between information extraction and summarization collapses. Consider the hypothetical summary in Figure 2.1, similar to the commitment summary example in Figure 1.1 (page 15) in the previous chapter. It is unclear whether this a summary or a unit of extracted information.

Summarization is also closely associated with **discourse analysis** (Sinclair and Coulthard, 1975; Brown and Yule, 1983), a term I shall use to indicate structural analysis of purpose and contextual relationships in language use. Components of discourse structure are often associated with large blocks of text or with far-reaching contextual relationships. This means that discourse analysis tasks such as *discourse segmentation* (Litman and Passonneau, 1993) or *rhetorical parsing* (Marcu, 2000) produce structures that might be called summaries, since they contain many fewer elements than the original text. A topic segmentation of a meeting with associated topics listed (Purver et al., 2006b), for example, can provide an overview of a meeting that is only a few lines in length. Or an argument diagram of a meeting (Rienks and Heylen, 2005; Verbree et al., 2006) can provide a rhetorical overview. Both constitute a kind of summary.

Because of the close relationship between summarization, information extraction, and discourse analysis, it is important to consider all three in our attempt to find a common set of descriptive concepts. This will allow us to evaluate the components of such systems comparatively, and to identify avenues for innovation that transcend their differences. For example, we may want to compare meeting summarization to determining the main arguments in a legal text (Grover et al., 2003), particularly since it is not unreasonable to assume that some meeting minutes should contain such argumentative information. For this reason, I shall be liberal with my application of Spärck Jones' framework, and I shall use it to describe summarization, information extraction, and discourse analysis systems collectively.

2.1.1 Spärck Jones' descriptive framework

System structure

As natural language information processing systems, summarization (and related) systems consist of a set of data models and processing stages which constitute the *system structure*. Following Spärck Jones (1999, 2007), this can be viewed as having three primary processing stages, depicted in Figure 2.2.

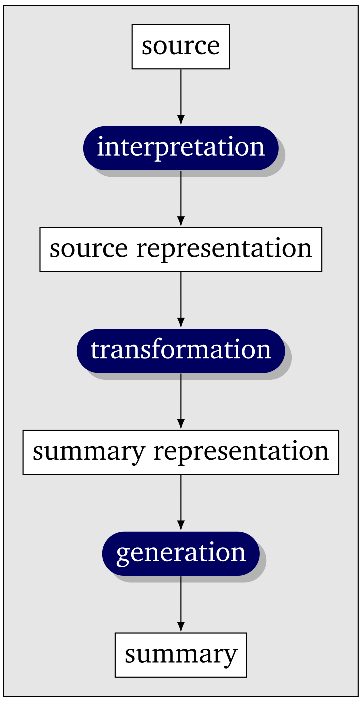
The **interpretation stage** takes the *source*, e.g., a text, database, or recorded conversation, and produces a *source representation*, typically by statistical and linguistic processing. A source representation might consist of a list of utterance identifiers with assigned relevance scores, e.g., Murray et al. (2005a), or it might be a discursive representation backed by rich domain knowledge, e.g., Alexandersson et al. (2000).

The **transformation stage** then creates a *summary representation* from the source representation. In the extractive summarization paradigm, transformation commonly involves ranking and selecting units from the source representation. Information extraction employs a similar approach, selecting units if they are examples of certain informational types, such as a membership relation between an individual and an organization, e.g., "Michael Smith is an employee of CyberInfo Corp." In the abstractive summarization paradigm, transformation embodies a more complex set of processes, usually involving some sort of further discourse-level interpretation.

The final stage is the **generation stage**, where a *summary* is created from the summary representation. Generation is typically use in more sophisticated summarization methods. For example, a system might pick out smaller elements of the source, e.g., at the phrase level (McKeown et al., 2002), and reconstruct sentences from them. Full natural language generation (i.e., starting purely from a logical form interpretation) has only been used in highly constrained domains in conversational speech summarization, due to the difficulty of producing such interpretations in open-domain settings (Alexandersson et al., 2000; Reithinger et al., 2000).

For applications centered around information extraction and discourse analysis, it is also unusual for generation to be used. Rather, the production of a structured data rep-

Figure 2.2: An information processing model of summarization systems (Spärck Jones (2007)).



resentation is often the ultimate goal in such applications. For example, an information extraction application might produce a table that lists any mentioned persons along with their occupations. A discourse analysis application might produce a graph structure depicting the arguments a speaker made, e.g., Rienks and Heylen (2005). In the meeting domain, summaries of action items or decisions blur this distinction between structured output and natural language output. Consider again the hypothetical decision summary in Figure 2.1 above. This summary is almost entirely encoded in a formal template-like structure, with a small amount of generated text used to complete the decision descriptions. Some summarization processes may therefore be considered akin to *template-filling* operations, with or without subsequent natural language generation (Purver et al., 2007).

Summarizing factors

The second part of Spärck-Jones' descriptive framework concerns *summarizing factors*, or characteristics that pertain to a summarizer's *context of use*. Spärck Jones (2007) identifies three main groups of such properties: *input factors*, *output factors*, and *purpose factors*.

Input factors relate to the system's input source, including its form, register, medium, and genre. Meeting summarization systems typically use at least the recorded speech as input, and most also use the output of automatic speech recognition as an additional input source. But systems can also supplement input with other sources such as video (Mikic et al., 2000) or handwritten notes (Banerjee and Rudnicky, 2009).

Output factors relate to the properties of a generated summary, thus encapsulating the commonly used distinction between extractive and abstractive summarization. An extractive meeting summary, for example, is typically composed of a list of transcribed utterances or speech spurts (Murray et al., 2005a). An abstractive meeting summary, on the other hand, will generate novel utterances that describe the conversation (Murray et al., 2010a,b). Output factors extend to many other properties of the output summary as well, including its coverage, style, format, and level of reduction. A decision detection application, for example, may provide output in the form of links to locations in the original media recordings (Hsueh and Moore, 2008).

The third group of factors are the **purpose factors**, which concern how the system and its summary are actually used. A summarization system that creates meeting minutes, for example, will likely employ purpose factors that directly relate to the traditional use of meeting minutes, such as for briefings and public dissemination (Whittaker et al., 2006). But other types of meeting summarizers might address entirely different uses, such as for the purpose of auditing previous decisions (Murray et al., 2009) or skimming quickly through a meeting transcript (Tucker and Whittaker, 2009).

System evaluation

The third main component of Spärck Jones' (2007) descriptive framework is *system evaluation*. Evaluation is a complex problem for summarization because each unique processing step, intermediate representation, and output is open to evaluation, as is the effectiveness of the system in its specified context of use. But rather than distinguish varying evaluation types as either *intrinsic* or *extrinsic* (Spärck Jones, 1996), as is commonly done, Spärck Jones (2007) presents a more nuanced analysis of an evaluation's relationship to task context.

Spärk-Jones observes that all evaluations are related to *intended purpose* in a variety of (sometimes unspecified) ways. Her categorization begins with **semi-purpose evaluation**, which refers to evaluations that appear to be completely independent of intended purpose, such as the evaluation of text quality (see Marcu and Gerber (2001) for discussion). Such evaluations are by definition intrinsic, but she points out that they are also related to purpose by virtue of there being an *assumed* relationship of some kind between the evaluated property and the usefulness of the summary. She suggests that this is important to understand and explain. For example, the popular ROUGE evaluation method (Lin, 2004) saves resources by proposing that n-gram overlap with a gold standard is a sufficient place-holder for a human evaluation. This kind of assumed correlation to the true task, however, has been shown to be highly susceptible to changes in summarization factors and uses (Murray et al., 2005b; Liu and Liu, 2010). A **quasi-purpose evaluation** also has an indirect relationship to purpose whereby an alternative task is evaluated that bears some assessed relationship to the actual task. For example, a pre-defined question-answering task might be used in lieu of users actually generating questions themselves. A **pseudo-purpose evaluation** takes one step closer to true purpose evaluation by *simulating* the true task context. The SUMMAC evaluations (Mani et al., 2002), for example, perform pseudo-purpose evaluation by employing models of analytical tasks such as relevance assessment. A **full-purpose evaluation**, on the other hand, does not use a model at all, but rather assesses a system directly within its fully-specified intended context of use.

2.2 Literature Review and Analysis

With a descriptive framework now in place, I shall begin a review of relevant work in automatic summarization, focusing on speech summarization (which now has a history going back almost two decades). The goal of this review is to identify the *dominant trends*, *general conclusions*, and *outstanding problems* in summarization research related to this thesis. This will provide further motivation and context for an activity-oriented approach.

2.2.1 The origin of speech summarization

Automatic summarization and automatic speech recognition have histories going back at least to the 1950's (see Luhn (1958) and Davies et al. (1952) for examples of incipient work in each field). But it was not until the late 1990s that summarization collided significantly with speech recognition, and the field of **speech summarization** emerged. Rapid developments in information extraction and retrieval were occurring at the time, as represented in the results of the Message Understanding Conference (MUC) (Grishman and Sundheim, 1996) and Text Retrieval Conference (TREC) series (Spärck Jones, 1995) of the early 1990s. Then in 1995, DARPA introduced broadcast news as a focus of their automatic speech recognition programs (see Pallett (2002) for a brief history), and very soon thereafter, the two research areas began to see significant interaction. In 1997, the spoken document retrieval task was introduced to the TREC conference (Garofolo et al., 1997). In the same year, the Topic Detection and Tracking (TDT) pilot study was conducted on transcripts of broadcast news (Allan et al., 1998). In 1998, several more important events took place. The DARPA/NIST broadcast news evaluations introduced a component on named entity recognition (Przybocki et al., 1999), and the TDT-2 evaluation added speech in addition to transcripts (Fiscus et al., 1998). These conferences and evaluations played a pivotal role in spawning the research area known as speech summarization (also referred to in the speech processing literature as spoken document summarization), and they set the scene for the vast majority of related research that has occurred in the ensuing 15 years.

2.2.2 Exploring speech-specific features

The initial stage of research on speech summarization was dominated by what may be called a 'speech-enhanced' approach, in which the results of *speech-specific processing* were incorporated into existing text summarization techniques. This was an obvious starting point, given that text-based methods were already well developed, but it prompted a focus that was decidedly on summarization *input factors*. Existing text-based approaches were amended to account for the new form of input that speech represented, and speech was in many respects viewed as a corrupted, unstructured stream of text requiring de-corruption and re-structuring. Some approaches also considered taking advantage of the supplemental evidence that speech provides, i.e., prosody and acoustics.

For the most part, speech-enhanced approaches employed the paradigm of content *extraction*, in which fragments of the original text are selected and reproduced in the summary. To get a sense of the problem and the solutions proposed by such methods, consider the human-annotated target summary shown in Figure 2.3, reproduced from Koumpis (2002). The example shows an automatic transcript of a voicemail, where some of the words (bold) have been selected for inclusion in a summary. This summary is designed to be extremely terse (for use on a mobile phone), and the coherence and flow of the output summary is not considered as important as the retrieval of the main 'topics' of the message.

Many of the first summarization systems to use speech-enhanced features focused on *broadcast news*. One of the earliest examples of this can be found in Valenza et al. (1999),

Figure 2.3: A human-annotated target summary for the voicemail domain (Koumpis (2002), p. 51). The original text is shown in its entirety in the first paragraph, and the target extract is shown in bold. In the second paragraph, the extract is presented on its own.

HI **BLAINE KAREN GATES** JUST WANT TO LET YOU KNOW I HAD TO **MOVE** THE **BIWEEKLY** WITH ASH- WITH **ASHOUK** AND **DRAGUTIN FROM JUNE THIRTIETH MONDAY TO TUESDAY JULY FIRST ELEVEN THIRTY** TO **TWELVE** THE SAME TIME BUT THE NEXT DAY UH I- IT WILL **NOT** HAPPEN ON THE **THIRTIETH** OF **JUNE** WE'RE GOING TO PUT IT **JULY FIRST** THANKS BYE BYE

BLAINE KAREN GATES MOVE BIWEEKLY ASHOUK DRAGUTIN FROM JUNE THIRTIETH MONDAY TO TUESDAY JULY FIRST ELEVEN THIRTY TWELVE NOT THIRTIETH JUNE JULY FIRST

Figure 2.4: A word-level spoken news summary produced by the system described in Hori et al. (2003a), p. 51.

Original sentence:	VICE PRESIDENT AL GORE SAYS THE GOVERNMENT HAS A PLAN TO AVOID THE INEVITABLE PROSPECT OF INCREASED AIRPLANE CRASHES AND FATALITY IS
70% summarization:	VICE PRESIDENT AL GORE SAYS THE GOVERNMENT HAS A PLAN TO AVOID THE INCREASED AIRPLANE CRASHES
40% summarization:	GORE THE GOVERNMENT HAS A PLAN TO AVOID THE INCREASED AIRPLANE CRASHES

whose major contribution was the use of acoustic confidence measures from speech recognition. Their approach was modelled on an information-extraction style of summarization, in which a confidence-weighted *tf-idf* score was used to identify the most relevant extracts. Another example of such work can be found in Hori et al. (2003b)[1], whose system extracts individual words rather than utterances to create a summary. Their approach introduced grammatical constraints on the output to accommodate word-level extraction. Like Valenza et al., they used an acoustic confidence score in word-selection. An example of the output produced by Hori et al. (2003a), p. 51 is shown in Figure 2.4[2]

Initial work on automatic summarization of *meetings* also investigated mainly speech-specific additions to text-based approaches. Waibel et al. (1998) presented the first application of summarization to meetings, for use within an integrated system for multimedia access and browsing (Schultz et al., 2001; Burger et al., 2002). Their approach used maxi-

[1] See Furui et al. (2004); Hori et al. (2003a) for further details and variations.

[2] Other approaches using recognition confidence scores have also been used for Mandarin news summarization (Hsieh et al., 2004; Wu et al., 2005; Huang et al., 2005).

Figure 2.5: An MMR-based utterance-level extractive summary of a telephone conversation (Zechner (2001b), p. 7).

B: They didn't know [...]
B: I mean it was a good place for the poor guy to die I mean because it was you know right after the rally and everything was on film and everything
B: Oh the whole country we just finished the thirty days mourning for him now you know it' uh oh everybody's still in shock it's
B: But what's the reaction in america really I mean I mean do people care you know I mean you know do they

mum marginal relevance (MMR) (Carbonell and Goldstein, 1998), applied to both manual and automatic meeting transcripts as an utterance selection method. The novelty of their method was to apply summarization to *turns at speaking* rather than sentences.

A more richly speech-enhanced approach arising from the same research group was later introduced in the DiaSumm system (Zechner, 2002). In DiaSumm, the interpretation and transformation stages of summarization were both a significant focus of development. Interpretation consisted of a pipeline of processing steps, including disfluency removal and sentence boundary detection. This was followed by a transformation stage in which sentences were selected according to MMR, but with additional consideration of their context within topical segments and question-answer pairs.[3] This inclusion of discourse-oriented features was a departure from previous methods and provided a considerable benefit to the quality of results. As an example, compare Figure 2.5 to Figure 2.6. The former presents the results of a pure utterance extraction method. The latter incorporates disfluency removal, sentence boundary detection, and the linking of question-answer pairs. (See Zechner (2001b), pp. 7–9 for the original conversation and further details about these examples.)

Classification-based methods The approaches of Zechner, Hori, and the others mentioned above, despite their unique enhancements, still relied centrally upon one main idea from text summarization—that the importance of a fragment can be based upon a statistical comparison of its content words to that of the larger corpus. One might call this an *unsupervised*, *distributional-statistical* methodology for sentence selection. However, another trend that was contemporary to this work took a computationally different perspective. This alternative speech-enhanced approach to summarization considered the extract

[3] See also Zechner (2001a,b); Zechner and Lavie (2001) for further details.

Figure 2.6: A summary of the same passage as 2.5, but with added disfluency removal, sentence detection, and question-answer linking (Zechner (2001b), p. 8–9).

B: It just worked out / it was a good place for the poor guy to die / because it was right after the rally / and everything was on film / the whole country we just finished the thirty days mourning for him now / everybody's still in shock
B: It's unbelievable / it's all those people you [...]
B: What's the reaction in america / do people care
A: Yeah / The jewish community all of us were very upset
A: Did it happen on a sunday
B: It happened on a saturday night

selection problem using a *supervised*, *classification-based* methodology, in which an extract classifier was learned from training data using machine learning techniques. One such system was a voicemail summarization system (Koumpis et al., 2001; Koumpis and Renals, 2005) that extracted speech-specific features such as pitch and duration as input to a word selection classifier. Improvements by introducing prosodic features alongside lexical features were also shown in other classification-oriented approaches to Japanese summarization (Inoue et al., 2004; Kobayashi et al., 2003; Togashi et al., 2006). And Kolluru et al. (2005) describe a system that performs disfluency removal and extract identification using a sequence of multi-layer perceptrons.

The supervised classification-based framework has remained an important approach, particularly with the advent of more accurate and powerful machine learning algorithms. The technique was first extended to the meeting domain by Murray et al. (2005a,b), and has been established as a powerful addition to the repertoire of speech summarization in general, particularly since it allows for straightforward integration of a diversity of information sources. A major problem with this approach, however, is that it requires a corpus of annotated examples on which to train the classifier.

2.2.3 Exploring discourse and interaction

Subsequent to the introduction of speech-specific approaches, another trend emerged in which improvements were sought through the use of discourse-level features, i.e., features that encode structure and relationships between sentences and groups of sentences. The initial work in this area was formulated by Zechner (2002) and Maskey and Hirschberg (2003), in which features relating to story segments and adjacency pairs were used (see McKeown et al. (2005) for a general discussion). Maskey and Hirschberg (2003), for example, identify genre-specific characteristics of newscasts to generate an outline which users can navigate interactively. The system assumes that the domain exhibits conventional discourse patterns from one document to the next, such as a sequence consisting of (1) an anchor's introduction, followed by (2) a news briefing, followed by (3) a reporters' in-depth presentation. The authors found that taking each of these steps into separate consideration provides an advantage over a document-wide thematic focus, since each step serves a rather different communicative purpose. Users expect different sorts of information to be extracted from these distinct parts.

Since initial work by Zechner, discourse-informed approaches have collectively studied a wide variety of features associated with both *discourse* and *interaction*. For example, Murray et al. (2006) and Murray (2008) investigate the contribution of turn-taking, speaker activity, and discourse cue phrases on meeting summarization. Zhu and Penn (2006) look at similar features for summarizing telephone conversations. Galley (2006) employs pragmatically-associated utterance pairs. Hirohata et al. (2005) select introduction and conclusion sentences. Fujii et al. (2007) look at the contribution of cue phrases. And Murray and Renals (2008) look at metacomments, i.e., utterances that reflect speech that is about the conversation itself. Finally, rhetorical structure has been employed for lecture summarization (Blair-Goldensohn and McKeown, 2006; Zhang et al., 2010). With such a wealth of discourse-level information being used, new questions have arisen: Which of these are most important to consider and when? What are their specific relationships to summarization? How is their importance dependent upon input, output, and purpose factors? These are some of the main questions currently facing the field.

Taking stock; Current directions So what general results can be gleaned from the first few years of speech-enhanced and discourse-enhanced approaches? One conclusion from the speech-enhanced approaches is that concentrating on the unique *input factors* of speech can pay off. Approaches that clean disfluencies, incorporate recognition confidence, or add prosodic features can be much more effective than simply using a speech recognizer's one-best output. However, most of the features that are thought to contribute have now been thoroughly studied, and significant advances relating to input factors appear to be waning.

The general results of discourse-enhanced approaches are mostly positive but difficult to interpret. Results generally point to the fact that specific genres of input demand rather different models and features, particularly as those features become more deeply linguistic.

For example, rhetorical structure and lexical features have been shown to work better for lectures than for news, while discourse segment features are better in news than lectures (Zhang et al., 2007). Adjacency pairs work for dialogue, but they are not applicable to texts or monologue. Another interesting result is that very simple discourse or interactional features, such as discourse cue words and sentence duration, are often very good predictors of sentence importance, even without any textual input at all (Maskey and Hirschberg, 2005; Murray et al., 2005a; Galley, 2006). And several systems have been developed where structural speech patterns were used successfully as a complete replacement for textual input (Maskey and Hirschberg, 2003, 2006; Zhang and Fung, 2007). This suggests that a full interpretation of discourse features may not be required, and that using shallow transformations, or even surrogates, of discourse features might be sufficient.

So what directions are the most recent investigations taking in light of these results? There appears to be a divergence occurring. On one hand, there remains a trend toward further development of the statistical underpinning of both supervised and unsupervised approaches described above. On this track, the most recent focus in on improvements in *pattern recognition*, i.e., the core statistical techniques being used. As mentioned above, feature development appears to have reached a plateau, and new avenues for signifiant advances now rest on *using* those features more effectively during the transformation phase rather than during interpretation. I shall not go deeply into this area, as the specific computational methods are not a major focus of this work.[4]

On the other hand, there is also a current trend toward summarizing different types of content, i.e., not just the important topics or expressions of subject matter. This is motivated by the knowledge that information about various kinds of discourse-level abstractions are generally more desirable to users (Murray et al., 2009). It has also been accelerated and by new advances in deep semantic and pragmatic linguistic processing. Techniques for deep understanding of discourse, however, have a strong dependence on genre, which makes them less flexible. Recent work therefore tends to focus on very specific types of speech with more structured interpretations. In the following section, I discuss

[4] Some approaches that fall into this category are (Chen and Chen, 2007; Chen et al., 2007; Lin et al., 2008; Chen and Chen, 2008; Chen et al., 2009; Lin et al., 2010), who incorporate a probabilistic generative model of topics. Another approach models extract importance in terms of concepts, focusing on the use of key phrases as a model of shared terminology (Riedhammer et al., 2008; Liu et al., 2008; Gillick et al., 2009; Xie et al., 2009; Riedhammer et al., 2010). Xie and Liu (2010) develop a new approach by considering extraction as a regression problem rather than classification. And sentence compression has continued to receive attention in the form of applying novel machine learning techniques (Garg et al., 2009).

examples of this kind of approach.

2.2.4 Discourse-focused summarization of dialogues

It is interesting to note that some of the earliest work on speech summarization took an orientation toward *discourse*. One example was BBN's summarizer for air traffic control dialogues (Rohlicek et al., 1992; Denenberg et al., 1993) which identified the location and type of sub-dialogues. Another was a conference room reservation dialogue summarizer developed at SRI (Kameyama and Arima, 1994; Kameyama et al., 1996) that used deep semantic processing to summarize the set of resulting reservations. There were also two systems that were contemporaries of Zechner's initial work on meetings. The CLARITY project (Finke et al., 1998; Levin et al., 1999; Ries, 1999, 2001a; Ries and Waibel, 2001) aimed to develop automatic classifiers for three levels of discourse structure in telephone conversations: speech acts, dialogue games, and discourse segments, as well as automatic labeling of speech acts. While not a summarization system *per se*, this program was an example of large-scale discourse analysis, the output of which can be seen as a kind of summary. Finally, the Verbmobil project (Wahlster, 2000) developed a translation system that contained components dedicated to summarizing negotiation dialogues (Alexandersson et al., 2000; Reithinger et al., 2000). This was a genre-specific system, where the summary relied on parsed sentences, logical forms, and discourse structures. Figure 2.7 shows an example of a human-annotated summary from the project (taken from Alexandersson (2003), p. 8).[5] This example clearly demonstrates the substantial difference between the discourse-oriented approach and the lexical and content-oriented approaches mentioned previously (e.g., Figures 2.3, 2.4, and 2.5).

Discourse-focused systems like VerbMobil are outliers in the history of speech summarization. These systems tend to consider summarization as more of an information extraction problem. They use rich semantic and pragmatic processing to perform discourse analysis during interpretation. And as their inputs are genre-specific, so too are their intermediate representations and output. This makes them highly dependent on (usually hand-coded) representations of the domain and discourse structure, which can take significant resources to implement. It also makes them problematic for spontaneous, unconstrained genres like workplace conversation. Alleviating this lack of generality is one of

[5] A complete consideration of relevant discourse analysis work within CLARITY may be found in Ries (2001b), and summarization work within Verbmobil is fully covered in Alexandersson (2003).

Figure 2.7: A translated human-annotated dialogue summary from the Verbmobil project (Alexandersson (2003), p. 8).

Participants: Speaker B, Thompson
Date: 6.8.2002
Time: 3:12 pm to 3:13 pm
Theme: Trip with accommodation and recreational activities

Result

Scheduling: Speaker B and Thompson agreed on a business meeting on the 20th of January 2002 at 11 am in Hanover. Speaker B and Thompson will meet on the 16th of January 2002 at 9:30 at the train station.
Traveling: A trip was agreed upon. The outward journey to Hanover by train starts at 5 pm on the 19th of January 2002.
Accommodation: A hotel in the city center was agreed upon. A single room costs 80 Euro. Thompson takes care of the reservation.
Entertainment: A dinner at a restaurant was agreed upon. Thompson will reserve a table.

the objectives of the work in this thesis.

Only very recently has a robust discourse-oriented method been attempted for continuous spoken language genres. Murray et al. (2010b,a) is the first to present such an alternative in a meeting summarizer. Their system looks very much like a collection of information extraction systems. It provides an integrated machine learning framework for identifying action items, decisions, and opinions, which are then collected and generated into a natural language summary. This work can therefore be seen as a descendent of specific information extraction applications from meetings, including action item extraction (Morgan et al., 2006; Purver et al., 2006a, 2007; Murray, 2008; Frampton et al., 2008), decision extraction (Hsueh and Moore, 2007a, 2008; Hsueh, 2008b; Fernández et al., 2008a,b; Bui et al., 2009; Frampton et al., 2009b), and agreement detection (Hillard et al., 2003; Galley et al., 2004; Hahn et al., 2006; Germesin and Wilson, 2009). And like these systems, it is a supervised machine learning system that is dependent on collections of human annotation. This questions its genre-dependence, and leaves open the question of whether a truly data-driven, unsupervised method is achievable in practice.

2.2.5 Purpose factors and information access applications

In her reviews of current trends in automatic summarization, Spärck Jones (1999, 2001, 2007) makes the suggestion that purpose factors are some of the most important factors that summarization researchers must address. She suggests that systems should be designed and evaluated in close correspondence to how they will be used. But this is a very difficult challenge for researchers to address, particularly given the fuzzy boundaries that exist between summarization, information extraction, discourse structuring, and other language processing applications. For example, a meeting browser system might decide to use headline generation, named entity extraction, sentence selection, and various forms of coarse-grained segmentation in its user interface. Must one then design a specific regime of evaluations suited precisely for this usage scenario? Or are there generic methods of design or evaluation that could be applied? Péry-Woodley and Scott (2006) suggest that this is a sign of things to come. They provide an overview of connections between discourse analysis and information access applications in which they set the scene for a messy convergence of discourse linguistics, document processing, and natural language processing—one that will be difficult to place within a common evaluation paradigm.

One line of research that addresses these concerns by analysis of purpose factors has been the work summarized in Whittaker et al. (2008). Whittaker et al. (2008) present a synthesis and evaluation of meeting summarization systems in which they compare them to related systems (Tucker and Whittaker, 2005, 2007), study the user requirements for such systems (Laban, 2004; Tucker et al., 2005), study the natural record-keeping behavior of meeting participants (Whittaker et al., 2006), and design a formal evaluation procedure for meeting browsers (Wellner et al., 2005; Tucker and Whittaker, 2009).

The results of Whittaker et al. (2008) provide a deep insight into current problems with meeting summarization. One of the main problems they identify is a lack of correspondence between what these systems present and how users naturally take notes and naturally want to use them. Firstly, they identify deeper semantic processing as an important avenue for progress. The systems they evaluated failed to provide the level of abstraction that would allow the user to strategically focus on the parts they deem important, such as agenda items, decisions, and actions. They also showed that browsers failed to address the contextual factors that are an essential complement to semantic content. Other evaluations of user requirements for meeting information systems generally confirm these conclusions (Conklin, 2003; Lisowska, 2003; Jaimes et al., 2004; Lisowska et al., 2004; Cremers and Hilhorst, 2005; Banerjee et al., 2005; Pallotta et al., 2006; Post et al., 2007; Ehlen et al., 2008), and point toward analysis of discursive notions such as arguments and opinions. They also suggest taking an orientation more closely geared toward participants and user needs.

2.3 Future Directions for Meeting Summarization

The literature review just presented reveals that automatic summarization of spoken conversations has been dominated by methods oriented toward extraction of important topics and subject matter. This solution generally does a good job of 'gisting' the conversation, but the unvarnished sampling of content-oriented words or utterances often makes the summary hard to read. The most recent research has addressed these problems by linking summaries to the original recording (Murray et al., 2009), generating headlines, or sentence compression (Zajic et al., 2008; Clarke and Lapata, 2010). But even with these enhancements, summarizing only subject matter makes these approaches fundamentally different to the natural summarization that humans do. Humans generally rely on a wide range of summarization mechanisms (Endres-Niggemeyer et al., 1995). They also tend to evaluate summaries more positively when they are based upon discourse-level features of the conversation (Marcu, 2000, p. 200), an approach more congruent with the approaches that were the outliers in our review above. Human summaries of conversations strongly favor the description of events, attitudes, and outcomes (Whittaker et al., 2006), not only concepts, facts, topics, and information otherwise representing the propositional content of the text or dialogue. The subject matter, or semantic content of a conversation is undeniably important to producing summaries. But participants' opinions, roles, decisions, and arguments are equally (if not more) important.

Because we need to reconsider what summaries of conversations really ought to be about, one of the main objectives of this thesis is to begin focusing on output and purpose factors rather than input factors. Steps toward such an objective have already been taken with respect to other settings of language use, particularly textual genres such as scientific articles (Teufel and Moens, 2002), legal texts (Grover et al., 2003), and email threads (Zajic et al., 2008; Murray et al., 2010a). Other novel work includes the detection of coarse activities in meetings from a vision-processing point of view (McCowan et al., 2005). But there is a pressing need to address the *discourse-oriented* aspects of the summarization problem as it relates to *spoken conversation*. Some incipient work has provided a starting point (Purver et al., 2007; Kleinbauer et al., 2007; Murray et al., 2010a,b), but a comprehensive, robust, data-driven approach is still needed. The study of meetings provides a diverse but coherent set of observations upon which to begin addressing this problem.

This work will not be a simple matter of improving current methods. It requires a wholly novel approach and a reconsideration of the specific problem context, i.e., the inputs, outputs, and purposes of summaries (Spärck Jones, 1999). We must understand how natural summaries of conversational speech are constructed, and we must study their relationship to elements of a conversation. We must address the problem of genre sensitivity, dependence on manual coding, and the robustness of deep processing. These are the initial prerequisites that will help to address the complex needs of users of integrated, interactive spoken conversation summarization systems.

Chapter 3

Communicative Activities and Conversational Summaries

In this chapter, I argue for an activity-oriented approach to summarizing conversations. By analyzing transcripts of workplace meetings and several kinds of human-produced summaries, I show how language use is a type of purposeful coordinated action, that commonsense conversational summaries focus on descriptions of communicative activity, and that such descriptions relate directly to participants' intentions. I also develop a template-like account of conversational activity descriptions, which I shall use to help design specific analytical tasks for subsequent experimentation. The emphasis here is on workplace meetings and meeting minutes, but many of my arguments are designed to address general problems in conversation summarization. I therefore introduce the Pear Stories corpus of conversational monologies as a means for testing generalization.

I also argue that communicative activities are principally indicated through participant-relational features, i.e., indicators of the relationship between participants and a conversation. I show how participant-relational features, such as subjective language, verbal reference to the participants, and the distribution of speech activity amongst the participants, can be used as a means for analyzing the nature and structure of episodic communicative activities. The goal is to develop analytical methods that are amenable to computation and which overcome the difficulties of the genre-dependent, taxonomy-driven approaches.

3.1 An Introduction to Activity-oriented Summarization

Conversations are more than just a linguistic product. They are examples of coordinated and purposeful *language use* (Clark, 1996). This elementary notion, a foundation for the approach taken in this thesis, may be usefully broken down into three fundamental approaches to language study: the intention approach, the action approach, and the interaction approach (all of which are encompassed by what Clark (1996, p. 56) calls the “action” approach). In the current study, using these approaches is not a matter of choice; it is necessitated by our problem. As I intend to show, they are essential to understanding commonsense summarization of conversation.

The *intention* approach to language use (Grice, 1969; Allen and Perrault, 1980; Grosz and Sidner, 1986) proposes that people engage in conversation for a reason. They use language as a means for achieving goals. Sometimes, the goals are clear from the outset. For example, when transacting a sale, a buyer wants to have something, and a seller wants to obtain money in return. Goals may be vague and unfold as a conversation progresses, as in a chance conversation between friends. Participants may also have different goals, such as a defendant wishing to be acquitted and a plaintiff seeking punishment. Whether gossiping or ordering a pizza, conversations don’t occur without participants having goals.

The *action* approach to language use (Austin, 1962; Clark, 1996) proposes that people use language as a form of action. They use language to do things. For example, people use language to sell and buy things, to gossip, to conduct trials, and to order pizzas. Language is not merely the exchange of information or facts. Utterances have an effect on the world. Participants use them to bring about change.

The *interaction* approach to language use (Sacks et al., 1974; Clark, 1996; Linell, 1998) proposes that language use is more than just action. Language use is a coordinated activity—an organized sequence of actions arising from the participation of ensembles of people. In other words, people use language *jointly* to achieve things *jointly* (Clark, 1996). For example, in a purchasing transaction, two individuals come together as seller and buyer. They organize their language into regular patterns of utterances. Each action contributes in sequence toward the achievement of their goals.

The combination of intention, action, and interaction is a natural one, with an intuitive interdependence among the three parts. Approaches to language with such a combination have been pursued in a number of disciplines, including artificial intelligence (Cohen and Perrault, 1979; Bratman, 1987), linguistics (Grosz and Sidner, 1986), psychology (Clark, 1996), and philosophy (Grice, 1957; Austin, 1962; Searle, 1969), work that has had a considerable and lasting influence upon the study of language in cognitive science (Stone, 2004). The current thesis is no exception, and I shall take it as uncontroversial that language use arises fundamentally as a result of intentional interaction. I shall henceforth refer to this as an **activity-oriented** approach to the study of language (see Allwood (2000) for a thoughtful methodological discussion).

In this particular study, it is important to take special note of intention, action, and interaction because they also happen to be elementary features of the *commonsense* view of language. In other words, they emerge in the way people (and most importantly, participants) *naturally* describe conversations. This is important, of course, because it relates directly to our current problem of determining what human conversational summaries are actually all about, and what automatically produced summaries should be about if they are to be natural and useful.

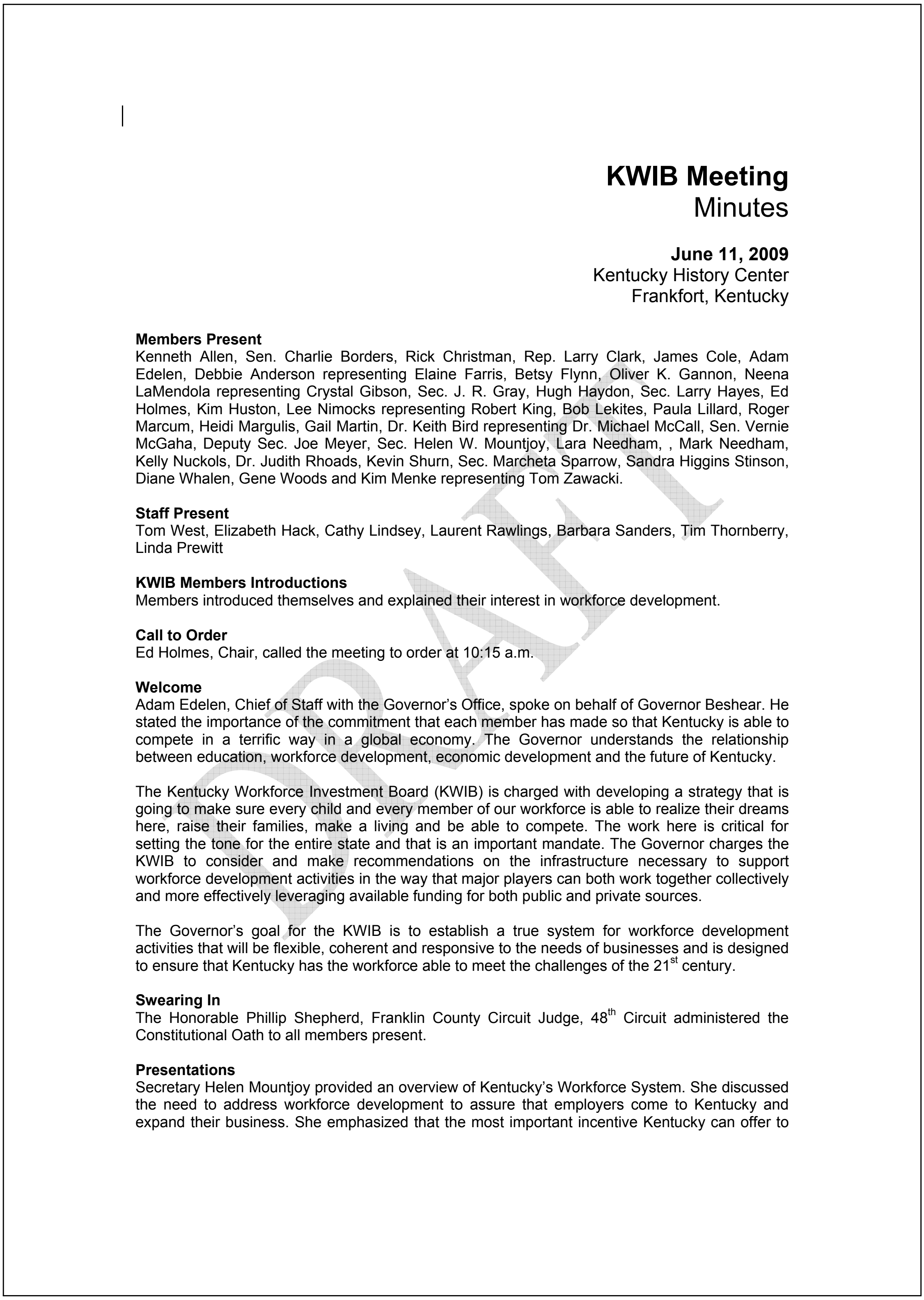
Clark (1996, p.33) writes about this in terms of a natural relationship between *goals*, *activities*, and *summaries*. He notes that joint activities, which he suggests are the underpinning of all language use, "can usually be summarized by describing the dominant goal achieved," for example, "The customer purchased two hotdogs from the street vendor" or "Police officer Clark interrogated the witness about a crime." These are invented examples, but Clark's idea brings up some basic notions that are worth exploring empirically. As it turns out, his account is remarkably accurate with respect to meeting minutes. Sentences like these abound in all types of conversational summaries. They are used by participants when summarizing their own conversations. They are used by participants within meetings themselves. They are used by annotators and minute-takers.

Consider, for example, the meeting minutes document shown in Figures 3.1 and 3.2, which contain the first two pages of the top-ranked result of a Google search for "meeting minutes filetype:pdf." A sketch of two basic properties of this document serves as an introduction to my basic arguments in support of an activity-oriented approach.

The first property is that section headings tend to describe *activities*—activities that the participants performed during the conversation, e.g., "Welcome", "Swearing In", "Presentations", "Visioning Exercise", and "Housekeeping." This trend is not universal. Some headings in this example also describe subject matter instead of an activity, particularly the sub-headings under "Board Actions," which include "KWIB Rules of Procedure" and "Incentive Grant Application." Others are headings indicating a list of members present. Nonetheless, the sequential listing of activities is a major feature of the high-level structure of the document.

The second property relates to leading sentences, such as "Secretary Helen Mountjoy provided an overview of Kentucky's Workforce System." These sentences also tend to describe the *activities* and *actions* of the participants. They often describe sub-activities of the

Figure 3.1: First page of the top-ranking meeting minutes PDF retrieved by Google.



KWIB Meeting
Minutes

June 11, 2009
Kentucky History Center
Frankfort, Kentucky

Members Present
Kenneth Allen, Sen. Charlie Borders, Rick Christman, Rep. Larry Clark, James Cole, Adam Edelen, Debbie Anderson representing Elaine Farris, Betsy Flynn, Oliver K. Gannon, Neena LaMendola representing Crystal Gibson, Sec. J. R. Gray, Hugh Haydon, Sec. Larry Hayes, Ed Holmes, Kim Huston, Lee Nimocks representing Robert King, Bob Lekites, Paula Lillard, Roger Marcum, Heidi Margulis, Gail Martin, Dr. Keith Bird representing Dr. Michael McCall, Sen. Vernie McGaha, Deputy Sec. Joe Meyer, Sec. Helen W. Mountjoy, Lara Needham, , Mark Needham, Kelly Nuckols, Dr. Judith Rhoads, Kevin Shurn, Sec. Marcheta Sparrow, Sandra Higgins Stinson, Diane Whalen, Gene Woods and Kim Menke representing Tom Zawacki.

Staff Present
Tom West, Elizabeth Hack, Cathy Lindsey, Laurent Rawlings, Barbara Sanders, Tim Thornberry, Linda Prewitt

KWIB Members Introductions
Members introduced themselves and explained their interest in workforce development.

Call to Order
Ed Holmes, Chair, called the meeting to order at 10:15 a.m.

Welcome
Adam Edelen, Chief of Staff with the Governor's Office, spoke on behalf of Governor Beshear. He stated the importance of the commitment that each member has made so that Kentucky is able to compete in a terrific way in a global economy. The Governor understands the relationship between education, workforce development, economic development and the future of Kentucky.

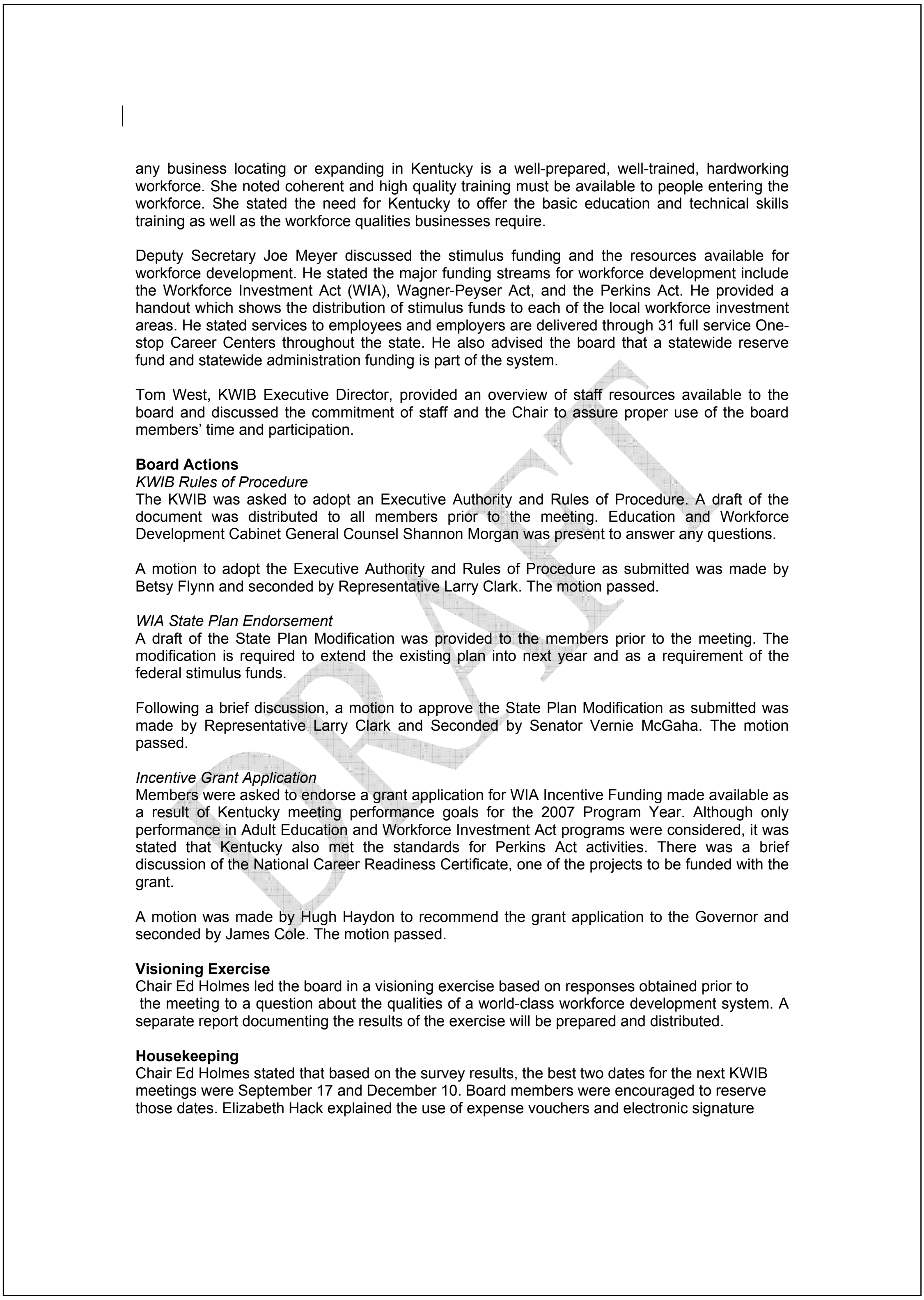
The Kentucky Workforce Investment Board (KWIB) is charged with developing a strategy that is going to make sure every child and every member of our workforce is able to realize their dreams here, raise their families, make a living and be able to compete. The work here is critical for setting the tone for the entire state and that is an important mandate. The Governor charges the KWIB to consider and make recommendations on the infrastructure necessary to support workforce development activities in the way that major players can both work together collectively and more effectively leveraging available funding for both public and private sources.

The Governor's goal for the KWIB is to establish a true system for workforce development activities that will be flexible, coherent and responsive to the needs of businesses and is designed to ensure that Kentucky has the workforce able to meet the challenges of the 21st century.

Swearing In
The Honorable Phillip Shepherd, Franklin County Circuit Judge, 48th Circuit administered the Constitutional Oath to all members present.

Presentations
Secretary Helen Mountjoy provided an overview of Kentucky's Workforce System. She discussed the need to address workforce development to assure that employers come to Kentucky and expand their business. She emphasized that the most important incentive Kentucky can offer to

Figure 3.2: Second page of the top-ranking meeting minutes PDF retrieved by Google.



any business locating or expanding in Kentucky is a well-prepared, well-trained, hardworking workforce. She noted coherent and high quality training must be available to people entering the workforce. She stated the need for Kentucky to offer the basic education and technical skills training as well as the workforce qualities businesses require.

Deputy Secretary Joe Meyer discussed the stimulus funding and the resources available for workforce development. He stated the major funding streams for workforce development include the Workforce Investment Act (WIA), Wagner-Peyser Act, and the Perkins Act. He provided a handout which shows the distribution of stimulus funds to each of the local workforce investment areas. He stated services to employees and employers are delivered through 31 full service One-stop Career Centers throughout the state. He also advised the board that a statewide reserve fund and statewide administration funding is part of the system.

Tom West, KWIB Executive Director, provided an overview of staff resources available to the board and discussed the commitment of staff and the Chair to assure proper use of the board members' time and participation.

Board Actions
KWIB Rules of Procedure
The KWIB was asked to adopt an Executive Authority and Rules of Procedure. A draft of the document was distributed to all members prior to the meeting. Education and Workforce Development Cabinet General Counsel Shannon Morgan was present to answer any questions.

A motion to adopt the Executive Authority and Rules of Procedure as submitted was made by Betsy Flynn and seconded by Representative Larry Clark. The motion passed.

WIA State Plan Endorsement
A draft of the State Plan Modification was provided to the members prior to the meeting. The modification is required to extend the existing plan into next year and as a requirement of the federal stimulus funds.

Following a brief discussion, a motion to approve the State Plan Modification as submitted was made by Representative Larry Clark and Seconded by Senator Vernie McGaha. The motion passed.

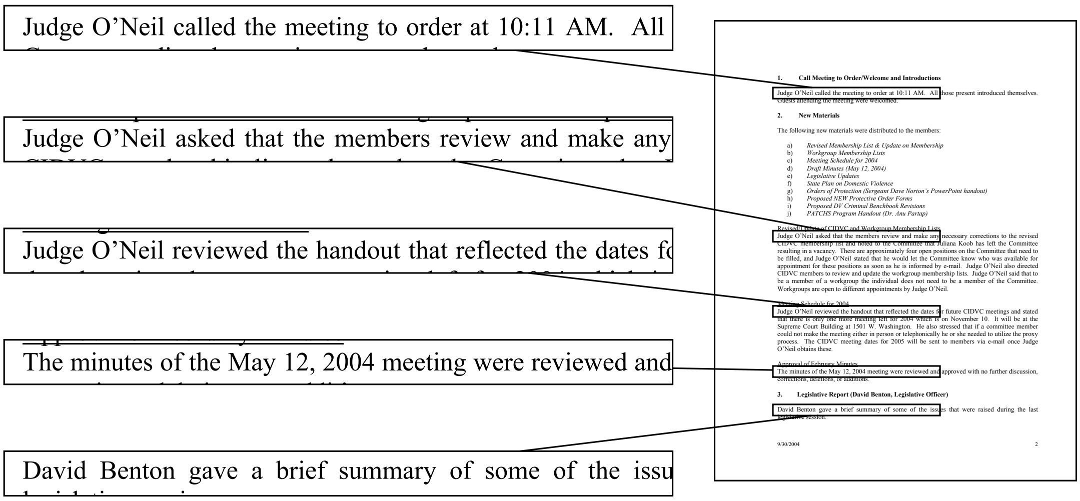
Incentive Grant Application
Members were asked to endorse a grant application for WIA Incentive Funding made available as a result of Kentucky meeting performance goals for the 2007 Program Year. Although only performance in Adult Education and Workforce Investment Act programs were considered, it was stated that Kentucky also met the standards for Perkins Act activities. There was a brief discussion of the National Career Readiness Certificate, one of the projects to be funded with the grant.

A motion was made by Hugh Haydon to recommend the grant application to the Governor and seconded by James Cole. The motion passed.

Visioning Exercise
Chair Ed Holmes led the board in a visioning exercise based on responses obtained prior to the meeting to a question about the qualities of a world-class workforce development system. A separate report documenting the results of the exercise will be prepared and distributed.

Housekeeping
Chair Ed Holmes stated that based on the survey results, the best two dates for the next KWIB meetings were September 17 and December 10. Board members were encouraged to reserve those dates. Elizabeth Hack explained the use of expense vouchers and electronic signature

Figure 3.3: Paragraph-leading sentences from a meeting minutes document. Example is from the August 2004 meeting of the Supreme Court of Arizona Committee on the Impact of Domestic Violence (publicly available).



Judge O'Neil called the meeting to order at 10:11 AM. All

Judge O'Neil asked that the members review and make any

Judge O'Neil reviewed the handout that reflected the dates fo

The minutes of the May 12, 2004 meeting were reviewed and

David Benton gave a brief summary of some of the issu

1. Call Meeting to Order/Welcome and Introductions

Judge O'Neil called the meeting to order at 10:11 AM. All those present introduced themselves. Guests attending the meeting were welcomed.

2. New Materials

The following new materials were distributed to the members:

a) *Revised Membership List & Update on Membership*
b) *Workgroup Membership Lists*
c) *Meeting Schedule for 2004*
d) *Draft Minutes (May 12, 2004)*
e) *Legislative Updates*
f) *State Plan on Domestic Violence*
g) *Orders of Protection (Sergeant Dave Norton's PowerPoint handout)*
h) *Proposed NEW Protective Order Forms*
i) *Proposed DV Criminal Benchbook Revisions*
j) *PATCHS Program Handout (Dr. Anu Partap)*

Revised/Update of CIDVC and Workgroup Membership Lists
Judge O'Neil asked that the members review and make any necessary corrections to the revised CIDVC membership list and noted to the Committee that Juliana Koob has left the Committee resulting in a vacancy. There are approximately four open positions on the Committee that need to be filled, and Judge O'Neil stated that he would let the Committee know who was available for appointment for these positions as soon as he is informed by e-mail. Judge O'Neil also directed CIDVC members to review and update the workgroup membership lists. Judge O'Neil said that to be a member of a workgroup the individual does not need to be a member of the Committee. Workgroups are open to different appointments by Judge O'Neil.

Meeting Schedule for 2004
Judge O'Neil reviewed the handout that reflected the dates for future CIDVC meetings and stated that there is only one more meeting left for 2004 which is on November 10. It will be at the Supreme Court Building at 1501 W. Washington. He also stressed that if a committee member could not make the meeting either in person or telephonically he or she needed to utilize the proxy process. The CIDVC meeting dates for 2005 will be sent to members via e-mail once Judge O'Neil obtains these.

Approval of February Minutes
The minutes of the May 12, 2004 meeting were reviewed and approved with no further discussion, corrections, deletions, or additions.

3. Legislative Report (David Benton, Legislative Officer)

David Benton gave a brief summary of some of the issues that were raised during the last legislative session.

9/30/2004 2

main activities, and they tend to present more details than the headers. One especially noticeable feature is that they have a typical linguistic pattern that involves a declarative event description in the past tense. This includes a participant or group of participants as the grammatical subject ("Secretary Helen Mountjoy"), an activity type description in the form of a past-tense verb phrase ("provided an overview"), and an attached description of the subject matter or topic ("of Kentucky's Workforce System"). Not all the sentences obey this simple template. Other sentences describe decisions, commitments, opinions, and supporting facts. Nonetheless, sentences describing activities and actions dominate as leading sentences for paragraphs, and they are common within the content of the paragraphs as well.

Consider, as another example, the extract of another minutes document in Figure 3.3. In this example, especially at the higher levels of structure, the document is dominated by simple past-tense declarative sentences describing the participants' activities.

The remainder of this chapter is devoted to filling in this basic sketch of an activity-oriented approach to conversational summarization. I shall explore the basic activity-oriented properties of natural conversational summaries. And I shall explore activities as they occur in conversations themselves, endeavoring to discover what relationships hold between the two. To understand how meeting minutes (and conversational summaries in general) are made, we must determine what types of information are included in natural summaries. We must understand how that information is presented. We must understand the elements of conversation that they describe. And we must determine how conversations are interpreted and reformulated into descriptions.

3.2 Summarizing Meetings: An Introduction to the AMI Corpus

I shall now expand on the nature of the summarization problem by way of a detailed example that explores a single conversation from many angles. The goal is to provide a deeper impression of how activities and summarization arise in the setting of multi-party workplace meetings.

The vast majority of analysis and experimentation in this thesis is based upon source data from a corpus of workplace meetings called the AMI Meeting Corpus (Carletta, 2007), which I shall now describe. The purpose of presenting this here is twofold: to describe in detail the dataset that will be used throughout the rest of the thesis, but also to give a general idea of what conversational summarization looks like in the setting of a workplace meeting. The AMI corpus is the most heavily annotated corpus of meetings in existence (Carletta, 2007), and we shall use these annotations as a valuable source of information for the current study.

Most of the meetings in the AMI corpus are the outcome of a designed meeting elicitation scenario where participants are experimental subjects and play the roles of employees in a fictitious electronics company. These will be referred to henceforth as **scenario meetings**. The subjects' task is to develop a prototype remote control device. Subjects participate in groups of four, and each participant in the group is assigned a distinct project role: project manager, industrial designer, marketing expert, and user interface designer. The group carries out the task over the course of four meetings during a single day. Each meeting is conducted according to a loosely-structured pre-defined agenda.

In the remainder of this section I shall introduce the scenario portion of the AMI corpus by example, providing qualitative empirical support for the importance of joint activities in conversational summarization. I shall do this by discussing five different types of source data available in the corpus: 1) speech transcripts, 2) summaries produced by the participants individually, 3) annotator-produced conversational summaries, 4) public meeting minutes produced by the project leader, and 5) annotator-produced topic outlines based upon a topical segmentation of the meeting. The presented examples all refer to the initial section of a single meeting in the corpus (ES2008A). All of the source data was produced as part of the development of the corpus, independently of my own investigations.

3.2.1 The dialogue transcript

The first source of information is the dialogue transcript itself, shown in example **(3)**. Looking first at the transcript (as opposed to annotations) allows us to focus initially on the participants' own natural, *commonsense* view of the joint activities in the conversation, since no annotation guidelines or other constraints have been imposed upon the participants' use of language within the meeting. While reading the transcript that follows, the reader is encouraged to try to identify the communicative activities that comprise sections of the dialogue, and the occasions when participants refer to these activities themselves (line 6 for example).

(3) [AMI–ES2008A–1]

1 D: Hmm.
2 A: Okay.
3 A: Good morning everybody.
4 A: Um I'm glad you could all come.
5 A: I'm really excited to start this team.
6 A: Um I'm just gonna have a little PowerPoint presentation for us, for our kick-off meeting.
7 A: My name is Rose [Anonymized].
8 A: I I'll be the Project Manager.
9 A: Um our agenda today is we are gonna do a little opening
10 A: and then I'm gonna talk a little bit about the project,
11 A: then we'll move into acquaintance such as getting to know each other a little bit, including a tool training exercise.[a]
12 A: And then we'll move into the project plan,
13 A: do a little discussion
14 A: and close,
15 A: since we only have twenty five minutes.

[a] The phrase "tool training exercise" refers to a drawing exercise in which the participants use a smart whiteboard to describe their favorite animal.

(3) [cont. . .]

16 A: First of all our project aim.

17 A: Um we are creating a new remote control which we have three goals about,

18 A: it needs to be original, trendy and user-friendly.

19 A: I'm hoping that we can all work together to achieve all three of those.

20 A: Um so we're gonna divide us up into three compa three parts.

21 A: First the functional design

22 A: which will be uh first we'll do individual work,

23 A: come into a meeting,

24 A: the conceptional design, individual work and a meeting,

25 A: and then the detailed design, individual work and a meeting.

26 A: So that we'll each be doing our own ideas

27 A: and then coming together

28 A: and um collaborating.

29 A: Okay,

30 A: we're gonna get to know each other a little bit.

31 A: So um,

32 A: what we're gonna do is start off with um let's start off with Amina.

33 A: Um Alima,

34 B: Alima.

35 A: sorry,

36 A: Alima.

37 A: Um we're gonna do a little tool training,

38 A: so we are gonna work with that whiteboard behind you.

39 A: Um introduce yourself,

40 A: um say one thing about yourself

41 A: and then draw your favourite animal

42 A: and tell us about it.

43 B: Okay.

44 B: Um I don't know which one of these I have to bring with me.

45 A: Probably both.

46 B: Right, so,

(3) [cont. . .]

47 B: I'm supposed to draw my favourite animal.
48 B: I have no drawing skills whatsoever.
49 B: But uh let's see, introduce myself.
50 B: My name is Alima [Anonymized].
51 B: Um I'm from the state of [Anonymized] in the US.
52 B: I'm doing nationalism studies,
53 B: blah, blah, blah,
54 B: and I have no artistic talents.
55 A: How do you spell your name?
56 B: A L I M A.
57 A: Thanks.
58 B: Oh,
59 B: and I guess I'm the Industrial Designer on this project.
60 B: So
61 B: let's see if I can get
62 B: um here.
63 B: I will draw a little turtle for you all.
64 B: Not necessarily 'cause it's my absolute favourite animal,
65 B: but just that I think they're drawable.
66 B: And you have the pretty little shell going on.
67 B: Some little eyes. Happy.
68 B: There you go.
69 B: That's a turtle.
70 D: Yes.
71 A: So what are your favourite characteristics?
72 B: Um. I I like the whole having a shell thing.
73 A: Mm.
74 B: It's quite cool carry your home around where you go,
75 B: um quite decorative little animals,
76 B: they can swim,
77 B: they can,
78 B: they're very adaptable,

(3) [cont. . .]
79 B: they carry everything they need with them,
80 B: um and they're easy to draw.
81 A: Excellent.
82 A: Shall we just go around the table?
83 C: Uh Okay.
84 C: Well,
85 C: my name is Iain uh
86 A: Mm.
87 C: and I'm the User Interface Designer for the project. Um.
88 C: And I'll try and draw my favourite animal.

One thing that is apparent about the dialogue in example **(3)** is that it contains distinct sequences of utterances that constitute conceptually unified blocks of discourse. For example, the first section [1–28] consists of the project manager introducing herself and the project. This might be subdivided into a greeting [2–6], a personal introduction [7–8], a summary of the agenda [9–15], and a description of the project [16–28]. The precise nature and type of these activities and their boundaries in the dialogue are fuzzy, and different individuals will likely discern somewhat different sections and descriptions. But overall, there is a generally discernible structure, identifiable as conceptually unified *episodes* of interaction.

So what meaningful concepts are the basis for describing these episodes? First and foremost, each episode is associated with a unique instance of a *type of communicative activity*. It is important to note that though most of the episodes involve only one person speaking, the activities being performed are undeniably *communicative*. And as such, they require that others are participating as addressees. Activity types in the example include greetings, personal introductions, summaries, and descriptions. These types of activities, of course, involve the participation of a group of individuals. (Clark (1996) refers to this collaborative aspect using the term *joint* activity.) But in addition to each episode centering upon a communicative activity, each activity in turn centers upon a particular topic, or subject matter. Communicative activities may thus be generally associated with a central *subject matter*. For example, the episode in lines 9–15 is a "summary" activity that is *about* an "agenda." The episode in lines 16–28 is a "description" activity that is *about* the "project."

My own descriptions of the episodes in the previous paragraphs, such as "introduction" and "project description," help to make a point. But a more convincing feature of the example is that the participants often refer to the activities themselves. The phrases "PowerPoint presentation" [6], "kick-off meeting" [6], "do a little opening" [9], and "talk. . . about the project" [10], are the first of many such references within the dialogue. Some of the references arise from the project manager's review of activities in the meeting agenda. Others arise from her description of the project in general. Others arise simply out of a natural tendency for participants to establish common ground about what they are doing (e.g., [47], [49], [63], [82], and [88]). These participant-produced expressions demonstrate that activity reference is part of the natural, *commonsense* appreciation that people have of conversation.

As a special case, consider the phrase "introduce yourself" on line [39]. Here, participant A is using the phrase as part of a request. What is special about this reference is that participant B subsequently fulfills the request during the ensuing dialogue [43–80]. Of course, just because a request is made does not always mean the request will be fulfilled. Nonetheless, this example shows that participants do make reference to their own conversational activities to help organize their interaction, a phenomenon that I will refer to as *bracketing meta-discourse* (Schiffrin, 1980). Such meta-discourse can be forward-looking, like this example, or it can refer back to a previous activity or achievement. With respect to analyzing a conversation, these special cases of activity reference are an important step toward validating an analysis in terms of participants' subjective experience of the conversation.

3.2.2 The participant-produced summaries

One can find further evidence of the significance of activities in participant-produced meeting summaries, shown in examples **(4)** through **(7)**. The following extracts are taken from summaries of the above conversation, produced by each of the participants at the end of the meeting. The participants were asked to do this as part of the experimentally-designed portion of the AMI corpus. They were given the following prompt: "Write one paragraph of coherent text to summarize the meeting as a whole from your role's perspective."

(4) [AMI–ES2008A–PartASumm]

1 This was a very introductory meeting.

(4) [cont…]

2 We focused mainly on getting to know each other and doing a little bit of brainstorming about the project.

3 It was the first time learning to use the tools, so we were all a bit apprehensive, but very willing to try.

4 I used a powerpoint presentation to give project aims, including finances, and then proceeded into a brainstorm about what we would each like in a television remote control.

5 Everyone participated fully and seemed engaged in the discussion.

6 A highly successful introductory meeting.

(5) [AMI–ES2008A–PartBSumm]

1 In this first meeting we discussed project goals, aims, finances, and the things to get done before next meeting.

2 We also took some time to become acquainted with the team members and the tools that we will be using.

3 To finish off the meeting we did some preliminary brainstorming on what the main features of our product design should be, based on previous personal experience with television remote controls.

4 For the next meeting I am to focus on the working design of the project, while others tackle other issues.

(6) [AMI–ES2008A–PartCSumm]

1 A 'kick off' meeting was held at 11 a.m. on 4th February 2005.

2 The meeting aimed to start a new project with the aim of designing a new television remote control.

3 After some introductions, an outline of the new project and the design process involved was given by the project manager.

4 General ideas were discussed, including the problems often encountered with existing remote controls.

(7) [AMI–ES2008A–PartDSumm]

1 At a bit after 11:00 on Friday 04 February, 2005, Rose, Alima, Ian and I (Jessie) met in the conference room to introduce ourselves, familiarize ourselves with the Real Reaction task and brainstorm ideas about the new television remote control.

(7) [cont. . .]

2 Rose, the project manager, led the meeting, and began by asking us to become comfortable with the dry-erase board by using it to draw our favorite animals.

3 Alima picked a turtle because she likes how it is able to travel with it's home, Ian drew a whale because he likes their mysteriosness, Rose illustrated a coyote like the ones she grew up listening to in California and I attempted to draw my favorite playful, water loving creature, a seal.

4 We then moved on to discussing ideas for the new remote, such as it's need to not be too complicated.

5 We got a bit sidetracked talking about remotes that would handle all sorts of electronics only to realize that this remote is solely to function the television.

6 We ended by dividing up tasks for the next meeting.

These extracts provide more evidence that sequences of activity are important to meeting summaries. In example **(4)**, one can find the phrases "getting to know each other" [2], "doing a little bit of brainstorming" [2], and "powerpoint presentation" [4]. The term *brainstorm* and its morphological variants can be found in three of the four summaries. The term *discuss* and its morphological variants occur in all four of the summaries. The *introduction* activity is referred to in each participant's summary, though in a variety of ways: "getting to know each other" [4:2], "become acquainted with the team members" [5:2], "some introductions" [6:3], and "introduce ourselves" [7:1]. These references to joint activities in the meeting provide evidence of their significance to participants' conception of what happened in the conversation. Their occurrence in multiple places, (i.e., in addition to mentions within the meeting) reinforces their importance.

3.2.3 The annotator-produced summary

As a slightly different source of summary information, consider example **(8)**, which is an *annotator*-produced summary of the meeting. Annotators were prompted by a fictitious email in which they were asked to summarize the meeting for a new project manager who would be taking over leadership of the project. The email instructed them to help this manager "stay informed about the state of the project," and to make notes that were "understandable for somebody who was not present during the meeting" (AMI Consortium, 2008b). For the first paragraph of the summary, they were asked to "write one paragraph of coherent text to summarize the meeting as a whole [...] think of this mostly as an abstract you would write for a paper (i.e., content-based), as opposed to a listing of sections (i.e., structure-based)." Example **(8)** shows the result of conducting this annotation task for the meeting currently under discussion.

(8) [AMI–ES2008A–AbsSumm]

1 The project manager opened the meeting and introduced herself to the team.

2 The project manager introduced the upcoming project in which the team is to create a remote control.

3 The team members participated in a tool training exercise in which they each drew their favorite animal on the white-board and discussed why they liked the animal.

4 The project manager then talked about the project finances and discussed selling prices, profit aim, market range, and production costs.

5 The project manager then led the team in a discussion on their experiences with remotes and what features they would like to include in the remote they are producing.

6 The team members discussed the option of combining remotes and how to produce a remote which is capable of controlling multiple devices.

In example **(8)**, one can find activity references similar to the participant-produced summaries, which again supports the importance of communicative activities in summarization. But looking more closely this time at how the activity descriptions are expressed linguistically can help to reveal some of their characteristics. For example, the annotator choses to use third-person past-tense verb phrases to describe most of the activities, including "opened the meeting," "introduced herself to the team," "introduced the upcoming project," "discussed why they liked the animal," and "talked about the project finances." Each of these verb phases have an associated subject noun phrase that refers to a participant or group of participants in the activity, e.g., "The project manager" and "the team members." This highlights the idea that they are activities performed *by the participants*, not simply events.

In some cases, the activity description is nominalized, e.g., "a tool training exercise" and "a discussion." Nominalizations appear to be used to provide nuance about the participants' involvement in it. For example, the verb phrases "participated in" and "led" indicate participants' roles and/or attitudes toward the activity. In addition, the movement of the activity type description from its usual spot in the verb phrase to an object noun phrase suggests that the activity plays a slightly less important role in these descriptions. Instead, the author is highlighting the *role* or *attitude* by expressing it in the verb phrase.

It is also interesting to note that due to the inclusion of activity descriptions, this summary may be considered both "structure-based" and "content-based," even though the annotation instructions requested that annotators focus on content. This could be due to the fact that annotators believed activities to be part of the "content" of the conversation, or that they found it difficult to describe content without placing it within a structure-oriented context.

3.2.4 The meeting minutes

The fourth source document is a meeting minutes document produced by the project manager. The first portion of the document is reproduced here as example **(9)**.

(9) [AMI–ES2008A–Minutes]

1 Minutes- Kick-Off Meeting - compiled by R. Surname
2 4/3/05
3 11:10
4 Agenda
5 - Opening
6 - Acquaintance
7 - Tool training
8 - Project plan
9 - Discussion
10 - Closing (we have 25 minutes!)
11 Project Aim
12 - New remote control
13 – Original
14 – Trendy
15 – User friendly
16 Project Method
17 - Functional design
18 – Individual work
19 – Meeting
20 - Conceptual design

(9) [cont. . .]

21 – Individual work

22 – Meeting

23 • Detailed design

24 – Individual work

25 – Meeting

26 Tool Training

27 • Try out whiteboard!

28 – Every participant should draw their favorite animal and sum up their favorite characteristics of that animal

29 Introductions

30 • Alima- Industrial Designer

31 – Drew a Turtle “pretty little shell”

32 – Nice to carry your home around

33 • Iain- User Interface Designer

34 – Whale- quiet intelligence, kind of mysterious

35 • Jessie- Marketing

36 – Seal- playful and silly

37 • Rose- Project Manager

38 – Coyote- sings to the moon, beautiful animal, association with home

39 Project Finance

40 • Selling price: 25 euro

41 • Profit aim: 50 M euro

42 – Market range: international

43 • Production costs: max. 12.50 euro

44 Discussion

[cont. . .]

In this meeting minutes document, we see mixed evidence of the importance of activities. There are a number of examples of activities, particularly at the highest levels of the hierarchical structure. For example, “Tool Training” [26], “Introductions” [29], and “Discussion” [44] describe activities. However, “Project Aim” [11], “Project Method” [16], and “Project Finance” [39] are more topical—they describe what the conversation was *about* rather than what the participants *did* or what they *achieved*.

The mix of activity and topically oriented top-level descriptions parallels what was seen in the minutes documents taken from the web. This particular example, however, does not seem to have many of the other typical characteristics that were observed in the web examples. The most striking example of this is that there are no complete sentences.

This is might be the result of the participants not having experience in minute taking, or it may have to do with the amount of time they were given to complete the document. The end result is a much more compressed summary of the meeting, with rather vague but concise descriptions of the conversational content. And it would seem that the extreme level of compression causes the author to adopt a focus on subject matter.

3.2.5 The topic outline

The fifth and final source of summary information in the corpus is a hierarchical topical outline of the meeting, created by an annotator who was instructed to segment the meeting hierarchically according to "what people were talking about—the topic—and when they changed topics" (Xu et al., 2005). (For each meeting, the topic segmentation and the summary were produced by the same annotator.) Annotators were also given a set of pre-defined topic labels that were expected to occur as part of the experimentally outlined scenario, though they were also given the freedom to assign novel topic labels if none of the pre-specified topics existed. Complete topic segmentation and labeling instructions to annotators may be found in Xu et al. (2005). Example **(10)** shows a textual representation of the topic segmentation and labeling annotations for the meeting under discussion (pre-defined topics are shown in italics). Each of these labels are associated with a temporally specified episode of dialogue.

(10) [AMI–ES2008A–TopicOutline]

- 1 *chitchat*
- 2 *opening*
- 3 project aim and goals
- 4 *drawing exercise*
- 5 *costing*
- 6 discussion of remote controls
  - 7 good and bad features
  - 8 batteries
  - 9 combining remote controls
  - 10 *look and usability*
- 11 *agenda and equipment issues*
- 12 *project specs and roles of participants*
- 13 *closing*

(10) | [cont. . .]

This very short episode-based summary of the meeting, like the minutes above, appears to mix topical headlines with activity-oriented descriptions. “Chit-chat,” “opening,” and “drawing exercise,” seem to convey activities, while “costing” and “batteries” are topic-oriented. This and the other summaries again support the notion that the main segments of a meeting are associated with either uniquely identifiable activities, or uniquely identifiable topics, or both. Again, it is interesting to note that the annotators were given some activity-oriented default descriptions, even though the instructions were explicit about the use of topics. This suggests that the term “topic” has a very flexible interpretation with some researchers and the general population. The interaction between these two dimensions will be discussed further below, and will be the specific subject of one of our experiments later in the thesis.

Section summary

In this section, I have presented empirical support from several sources for the idea that conversational summaries contain descriptions of the communicative *activities* of participants, i.e., the things that people *do* in conversations. I have given some examples of how activities occur as *episodes* within conversations, and I have shown that most activity descriptions are also associated with a particular *subject matter*. The descriptions showed a consistency across the different source types, both participant-produced and annotator-produced, which shows that the observed trends are part of our universal *commonsense* view of what conversations are all about.

Activity descriptions are, of course, not the only kinds of descriptions in conversational summaries. Many other kinds of descriptions also appear, such as topical headlines, expressions of opinion, cited facts, or documentation of future commitments. Some of these appeared in the summaries above. My approach in this thesis is to leave analysis of these other types of descriptions for future work. Rather, in the remainder of this thesis, I shall concentrate my effort on learning how to produce activity summaries, having shown that these are an essential ingredient of a natural conversational summary, particularly at coarser levels of analysis.

I would now like to dig deeper into the problem of communicative activity summaries and arrive at some more specific claims about the nature and structure of activity descriptions. I divide this presentation into two sections. Section 3.3 deals with activity *descriptions*. Section 3.4 deals with the communicative processes, i.e., the *discourse*, they describe.

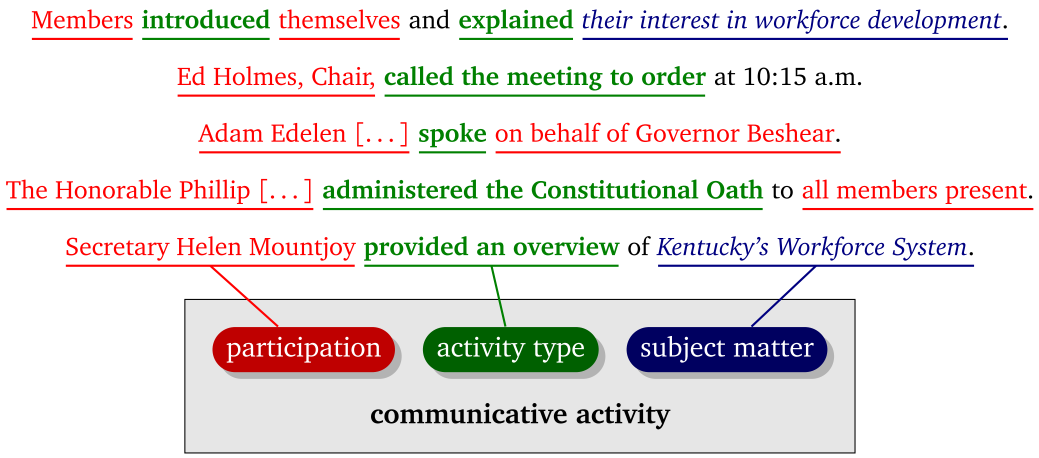
3.3 Activity Descriptions

The previous section showed how meeting minutes and other types of conversational summaries often contain **activity descriptions**. In their typical form, these descriptions occur as sentences that describe a single, episodic communicative activity occurring within a conversation. The purpose of this section is to present claims about what these individual activity descriptions *mean* and how they are semantically and syntactically constructed.

I begin by suggesting a prototypical structure in which the syntactic parts of activity descriptions are associated with three main semantic components, as shown in Figure 3.4. (The examples are paragraph-leading sentences taken from the minutes document in Figure 3.1.) The structure is comprised of three components: participation, activity type, and subject matter, which I henceforth abbreviate as **PAS**. The first component is the *participation* component, which provides meanings relating to who is performing the activity and in what capacity. This typically involves the use of a reference to a person or group as the grammatical subject of the sentence, usually in the form of a proper name, title, or role description. The *activity type* component typically involves the use of a communicative verb phrase, which describes the activity as an instance of a type of interaction. In some cases an activity type description is minimally informative, e.g., “spoke.” In others, it is more detailed, e.g., “administered the constitutional oath.” The third component is *subject matter*, which provides meanings that concern the topic of the activity, i.e., what is being talked about.

Note that these components are the *semantic* parts of a *prototypical* structure—they are the basic elements of meaning in the most common forms of activity description. The syntactic features also suggest typical, rather than required, usage. In the remainder of this section, I shall develop the details of the PAS components and explain how they interact with each other.

Figure 3.4: The three basic semantic components of activity descriptions: *participation*, *activity type*, and *subject matter*.


Members introduced themselves and explained their interest in workforce development.
Ed Holmes, Chair, called the meeting to order at 10:15 a.m.
Adam Edelen [. . .] spoke on behalf of Governor Beshear.
The Honorable Phillip [. . .] administered the Constitutional Oath to all members present.
Secretary Helen Mountjoy provided an overview of Kentucky's Workforce System.
participation
activity type
subject matter
communicative activity

3.3.1 Activity types

Activity description depends on an appreciation of commonsense, lexically-based categories of interaction. This is because activity descriptions tend to be declarative sentences containing verbs that denote the *type of activity* that took place. To explore the nature of these categories quantitatively, I shall examine the occurrence of verbs in a corpus containing the top 100 results of a Google query for "meeting minutes filetype:pdf" (the same query referred to earlier in this chapter). Employing a part-of-speech tagger (Toutanova et al., 2003), 516 unique past tense verb forms were automatically identified in this corpus. Table 3.1 contains a list of the most frequently-occurring verbs, ranked according to number of occurrences.

This list provides evidence that descriptions of communicative activities are central to meeting minutes. Almost all of the verbs denote types of *speech acts* (Searle, 1969) or can be categorized as *verbs of communication* (Biber et al., 1999). Of course, this is a rather intuitive outcome. It is hard to imagine that other kinds of events would have the same relevance to publicly-documented meetings. In genres of conversation where the physical world is more central, such as a task-oriented dialogue involving the manipulation of a physical artifact, one would expect to see non-communicative activities playing a more prominent role. However, for this small corpus of public minutes, which contains mainly minutes from legal, medical, governmental, and academic organizations, the evidence is in favor of communicative purposes being most important.

Table 3.1: Occurrences of the 30 most frequent past tense verb forms in 100 meeting minutes documents retrieved from the World Wide Web. Primary and light verbs are italicized.

1761	*was*	**108**	discussed	**57**	received
755	*were*	***77***	*made*	**57**	presented
596	*had*	**76**	advised	**57**	informed
347	said	**73**	explained	**51**	included
292	reported	**72**	requested	**51**	felt
235	asked	**70**	introduced	***51***	*did*
194	agreed	**64**	seconded	**47**	proposed
181	suggested	**64**	provided	**45**	indicated
180	stated	***63***	*gave*	**45**	approved
140	noted	**58**	thanked	**44**	reviewed

Another interesting feature of the list in Table 3.1 is that some of the words indicate events that are unlikely to have been realized as a single utterance. Rather, they refer to events one would expect to be realized as extended interactions, e.g., *discussed*, *explained*, *presented*, and *reviewed*. Reference to such extended events becomes more apparent as one explores the tail of the list not shown in the figure, which includes words like *recommended*, *reminded*, *decided*, and *outlined*. Most of these words, of course, are ambiguous about the duration of the events they describe. The events could have been realized as individual utterances or as sequences of them.

While communicative activity verbs dominate the list, other important classes of verbs also occur. There are, for example, verbs that do not (necessarily) denote *verbal* activities at all, e.g., *provided* and *received*. These are better characterized as denoting types of *interaction*. There are also several *cognitive* verbs, e.g., *felt*. Still, it is clear that the majority of terms denote communicative activities.

It should also be noted that primary verbs (Biber, 1999), i.e., *be*, *have*, and *do*, as well as the 'light verbs' (Butt, 2003) *make* and *give* are shown in grey despite being the most frequent. These verbs do not tell us much about the semantics of the sentences in which they occur, because they are typically attached to other words which carry most of the information. Further analysis of these words in the corpus shows that they are often used in phrases such as "gave a presentation," "made a comment," and "had a discussion." In these phrases, information about the activity type is expressed mainly in the verb phrase's nominal object, but it is usually straightforward to paraphrase them as verb-centered activity descriptions, e.g., “presented,”, “commented,” and “discussed” respectively.

3.3.2 Participation

Activities are more than simply events. They have participants who bring them about. Activity descriptions therefore often express factors of *participation*, such as who the participants in the activity were, what *roles* they played, or even their *attitudes* toward an activity. In prototypical PAS descriptions, this comes in the form of a person-referring expression as a sentential subject. What such a construction typically denotes is that the referenced person or group is the agent(s) of the activity, e.g., “*Judge O’Neal* reviewed the handout.” But the nature of participation can be richer than this, and it commonly involves more than just nominal subjects and expression of agency. Participation is a sententially expressed propositional notion, and the details come out in the interaction between a person-referring expression and the activity verb. To dig deeper into this issue, it is helpful to turn to a resource such as VerbNet (Schuler, 2005).

VerbNet is an inventory of verb classes, based upon a classification developed in Levin (1993). In VerbNet, each verb (lemma) has different senses that are in turn assigned to a verb class. Each class contains a set of frames, which are essentially common grammatical configurations (with associated interpretations). Each frame is then assigned a set of thematic roles and selectional restrictions for its arguments.

It is interesting to take note of the relationship between our list of common meeting minutes verbs and the verb groupings in VerbNet. By automatically mapping the first 100 verbs in our list into their appropriate VerbNet class (when available, and leaving out primary and light verbs), 82% of the mappable verbs fall into one of only two main groups of verb classes (covering those numbered 36 and 37). The first category contains the agent-based communication verbs (with an optional recipient). The second category contains the symmetrical communicative verbs, where participation is equal amongst all participants.

To exemplify the first category, consider the three most common activity verbs in our corpus: *said*, *reported*, and *asked*. VerbNet assigns a total of three possible thematic roles across all possible frames for *said* (SAY-37.7): AGENT (+animate, +organization), TOPIC (+communication), and RECIPIENT (+animate,+organization), with the RECIPIENT role being optional in six of eight frames. For example, the frame exemplified by the sentence

"Ellen said a few words to Helen" is assigned the syntax AGENT V TOPIC {TO} RECIPIENT. The same set of frames is provided for the verb *report*. (*Report* is actually considered a member of the same verb class as *said*.) And the same overall allocation of thematic roles is given for *ask* as well, i.e., INQUIRE-37.1.2.

To exemplify the second 'symmetrical' category we turn to verbs such as *debate*, *chat*, and *argue*. These verbs denote activities with multiple participants that each have an equal role in the activity. For these verbs, VerbNet uses the ACTOR role, and this role is shared across all participants. VerbNet uses this configuration for many communication classes, such as CHITCHAT-37.6 and MEET-36.2.

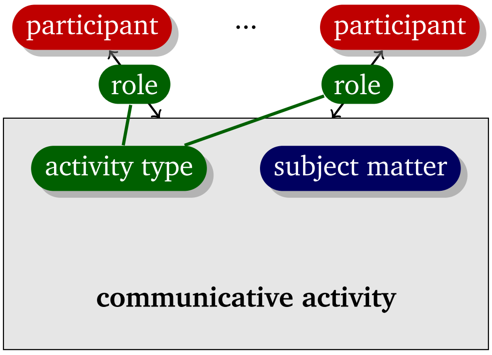
This quick look at VerbNet shows that the semantics of activity types and person-referring play a *joint* role in the expression of participation in activity descriptions. In other words, a *framework* of participation (or role structure) is made available by each activity type. A complete (propositional) activity description then *instantiates* the activity type, placing specific individuals into a particular configuration of those roles.

A summarization system must therefore know how to map observed behavior into these configurations of roles, and it must understand how different verbs can express them. Fortunately, mapping them to VerbNet shows that the most commonly observed verbs conform to a very limited set of role configurations, a fact that is likely to play a key role in making activity summarization more tractable.

Activity descriptions can, of course, elaborate upon more nuanced kinds of participation than that described above. Certainly, one can usually think up roles and relationships that go well beyond agent, recipient, and actor. Participants can take up all kinds of roles in interaction, such as the *proponent* of an idea or the *leader* of a discussion. But it will not be within the scope of this thesis to develop a full explanation of the likely complex architecture of participation in activities (though the simple model that emerges from the classes in VerbNet appears to be an effective direction toward making simplifying assumptions).

An important point that *is* within scope for this thesis, however, is the idea that activity type and participation may be seen as two sides of the same coin. The basics of this idea are summarized graphically as in Figure 3.5, in which the three PAS components are placed into a relational structure. The diagram shows how individuals play different roles as participants in an activity. Note, however, that specific roles found in VerbNet are not realized in the diagram. Rather, the roles are represented *generically* in terms of their association to the activity type. What this association is intended to imply is that there

Figure 3.5: An elaboration of the PAS components, introducing separate roles for participants and the direct association between activity type and role (participant-relational) structure.



is a direct correspondence between activity types and role configurations. For example, a lecture involves (typically) one lecturer and several students. But a lecture does not just *involve* participants in these roles, it is (partly) *defined* by them. Being a lecturer in interaction with a group of students is part of what it means to be 'lecturing'. This structured, relational representation of the meaning of a communicative activity will be expanded upon later, where I will generalize this further, replacing *role* with the more general notion of a *participant relation*. This structure is at the very heart of the participant-relational approach taken in this thesis.

3.3.3 Subject matter

The third component of activity description is the topic or *subject matter* of a communicative activity. Being communicative, conversational activities naturally have something that their participants are talking *about*. For example, reports are given *about* events or facts, or decisions are made *about* future action. This makes subject matter the most flexible and abstract component to the PAS structure. It can include facts, propositions, things, future events, and attributes, among many others. In the two VerbNet categories described in the previous section, roles for subject matter are given the names THEME, TOPIC, or PREDICATE.

The most important point that the PAS structure makes with respect to subject matter is that subject matter is *not the central component* of a conversational description. Rather, it is an element of a greater structure centered on activities. This contrasts with usual approaches to summarization, which focus almost entirely on subject matter, employing concepts such as topic, lexical cohesion, and reference as the dominant means for understanding conversation. Instead, the PAS structure makes the point that subject matter is a *complement* to activity type and participation.

But beyond this basic statement, further detail for the subject matter component shall not be drawn out. There appears no reason to suggest that traditional topically-oriented models of conversational meaning need to be changed to suit the activity-oriented approach proposed in this thesis. Rather, what I shall hope to demonstrate later on is that a semantically-focused analysis and an activity-oriented analysis are not mutually exclusive. Rather, they are mutually beneficial.

Section summary

In this section, I have identified three main semantic components of activity descriptions: *participation*, *activity type*, and *subject matter*. A simple quantitative analysis of past tense verbs in a corpus of meeting minutes demonstrated that the majority of such verbs describe communicative activities. A qualitative assessment of the verbs also suggests that many of the activities being described are likely to have been realized as extended periods of talk, i.e., *discuss* and *present*. And by looking at how VerbNet represents the same verbs, it was found that they fall into two main groups of verb classes, those with an agent-recipient role structure, and those with a symmetric actor-actor role structure. Together with a description of what is being talked *about*, this role structure and the category of interaction denoted by the verb itself are the principal three components of a prototypical communicative activity description.

3.4 Activity Summarization as Discourse Analysis

The problem I shall now address is the *analysis* of joint activities in conversation, i.e., the *interpretation* step (Spärck Jones, 1999) in an activity-oriented summarization system. In other words, I shall explore the question of how a summarization system, through observation of a conversation, might arrive at a set of activity descriptions like those just discussed. Where the previous sections have discussed the products of such an analysis, in this section I shall now investigate the essential analytical tasks that must be performed:

Task 1 Locate the main communicative activities within the dialogue.

Task 2 Recognize each activity's type, subject matter, and participation structure.

Task 3 Create an activity-centered natural language description for each activity.

My method for tackling these problems in the current section is to conduct a critical review of relevant approaches in discourse analysis. The section's main contribution shall be to identify the most plausible approaches from those in the current repertoire. This will involve studying unresolved issues, identifying the roadblocks that current methods face, and coming up with a set of requirements for progress.

3.4.1 Discourse interpretation: A critical review

The analytical requirements of an activity-oriented summarizer contrast significantly with the requirements for a traditional, subject matter-oriented one. The main distinction is that for activity-oriented summarization there is a requirement that the process be connected to *discourse interpretation*. Activity-oriented summarization depends upon providing a coherent interpretation at larger levels of the linguistic structure. Indeed, this is a basic hypothesis underlying some prominent discourse theories. For example, Mann and Thompson (1988) propose that the root (or nucleus) of discourse structures have a direct correspondence to appropriate summaries for their spans. But the structures defined within various discourse models contribute different perspectives on what is important. The Grosz and Sidner (1986) theory of discourse structure, for example, expresses speaker purposes. Rhetorical Structure Theory (Mann and Thompson, 1988), on the other hand, expresses rhetorical relations. The usefulness of each approach therefore depends on the input, output, and use factors of the summarization system (Spärck Jones, 2007). At the same time, it is unlikely that any single representational theory can capture all the information necessary for a good summary in any domain (Spärck Jones, 1993).

The complexity of discourse interpretation

A basic issue, and one that has likely been a significant reason for slow progress on discourse-oriented summarization, is that interpreting a discourse is an extremely complex, multi-dimensional problem. Consider, for example, one approach to the problem—that of producing an intentionally-oriented interpretation of a dialogue. First of all, there are a great number of different types of activities that participants perform across the many arenas of language use (Levinson, 1992). Participants also usually perform activities in pursuit of several goals at any given time (Hobbs and Evans, 1980), some of which may be shared amongst the participants while others may be held individually (Grosz and Kraus, 1996, 1999). The goals themselves, as well as participants' plans for achieving them, change regularly and spontaneously throughout the activity (Grosz and Hunsberger, 2006). And their achievement requires participants to perform complex, extended sequences of co-ordinated actions (Sacks et al., 1974). These individual actions, considered as units of language use, themselves perform multiple functions (Bunt, 2006), building and advancing a dialogue. But in most cases, functions are not explicitly expressed, but are instead implicit from various forms of context, including the attentional state of the participants (Grosz and Sidner, 1986; Grosz et al., 1995), their shared (i.e., grounded) knowledge of the situation (Clark and Brennan, 1991; Traum, 1994), and norms of interaction and co-operation (Grice, 1975). To add to this, the complexity involved is particularly apparent in natural human-human multi-party conversations (Traum, 2004), e.g., the language of face-to-face meetings. Interruptions, disfluencies, asides, grounding problems, overlapping speech, and a generally spontaneous and fluid character all contribute to the difficulty of analysis. To be able to model the intentional factors of spontaneous conversation thus represents a significant, far-reaching problem, one whose solution will require a great many interacting components of a diverse nature. And this list of issues does not even reflect any of the issues in more rhetorically-oriented paradigms of discourse interpretation. It is therefore important that our chosen approach settle on a robust, greatly simplified model of activities if it is to succeed.

Computational approaches to discourse interpretation

The computational ideas on which our approach might be based fall within the scope of the field of *computational pragmatics* (Bunt and Black, 2000; Jurafsky, 2005), which is concerned generally with the development of formal, systematic models that describe language and its relationship to action, reasoning, purpose, place, time, and other 'contexts.' There are four main problems this field has addressed: reference resolution and generation, interpretation and generation of speech acts, interpretation and generation of discourse structure and discourse relations, and abduction (Jurafsky, 2005). While it is impractical to entirely separate these problems from one another, for the moment it shall be helpful to focus on *discourse structure* in order to survey the main thrusts of research in computational pragmatics, since theories of discourse structure often underly approaches to other problems, e.g., reference resolution.

Computational models of discourse structure may be distinguished into two general categories—those that focus on cognition and those that focus on information. The cognitively-oriented approaches tend to concentrate on the role of *attention* and *intentionality* in linguistic processes (Grosz and Sidner, 1986; Grosz and Kraus, 1996; Lochbaum, 1998; Larsson, 2000), and have significant ties to work in speech act theory (Austin, 1962; Searle, 1969, 1976), cognitive psychology (Clark, 1996), cognitive science (Reichman, 1978), and artificial intelligence (Cohen and Perrault, 1979; Allen and Perrault, 1980; Reichman, 1985; Bratman, 1987). The informationally-oriented approaches tend to concentrate instead on the *rhetorical* and *linguistic* aspects of discourse, i.e., the interface between discourse, grammar, and meaning. For example, Rhetorical Structure Theory (Mann and Thompson, 1988; Taboada and Mann, 2006) addresses text organization by means of a hierarchical structure in which components have rhetorical roles and relations to one another, such as *elaboration* or *circumstance*.

Moser and Moore (1996) propose that rhetorical (informational) and intentional (cognitive) theories of discourse structure can be viewed as two sides of the same coin, and some work has explicitly endeavored to *combine* the cognitive and linguistic view (Stone, 2004; Stone and Lascarides, 2010). However, because of the intimate connection between intentionality and the notion of *activity* being focused upon in this thesis (see previous discussion in Section 3.1), it is the cognitively-oriented work that will be more directly relevant to our current problem. I shall therefore now focus on discussing some of this work in more detail, looking at how intentional characterizations of discourse have been applied to naturally occurring spoken conversations.

Intentional discourse analysis and dialogue coding

In recent years, large-scale corpus annotation studies have been particularly effective at advancing computational discourse research. Some recent examples include studies of anaphora in task-oriented dialogues (Poesio, 2004b), discourse relations in the Penn Treebank (Prasad et al., 2008), and dialogue acts in workplace meetings (Shriberg et al., 2004). These studies have provided a means for testing specific hypotheses, testing the reliability of human annotation, and training computational models that can perform analysis automatically. Studies that have investigated natural, multi-party, face-to-face conversation are the most pertinent to our current problem, and within that area of research one of the most common types of corpus studies have been those looking at the *functional* properties of discourse units, i.e., the goals, purposes, actions, and intentions embodied by those units. The generic label for this kind of research activity is *dialogue coding*.

Most of the literature on dialogue coding has focused on the labeling of *dialogue acts*, which can be summarized as the assignment of functional (often illocutionary) categories on an approximately utterance-by-utterance level (terminology and segmentation varies depending on the theory informing the task, see Bunt et al. (2010)). This has a clear relevance to this thesis, since the activities we are interested in are also based in a functional analysis of the dialogue—like the annotation of speech acts, the annotation of communicative activities is grounded in an intentional, goal-oriented characterization of language use.

Consider some of the categories used in previous work on coding of task-oriented dialogues. Heeman (1993), for example, coded the TRAINS corpus using four labels: *suggest*, *request*, *accept*, and *reject*. These labels were not designed to capture all possible utterance functions, but were instead designed to account for changes in a plan-based model of their task-oriented dialogue scenario. The MapTask scheme (Carletta et al., 1996, 1997) embodied a slightly more complex structure, again designed for task-oriented dialogue. It contained a hierarchical typology of speech acts rooted principally in the abstract categories of *initiate* and *respond*. The Verbmobil scheme (Alexandersson et al., 1998; Wahlster, 2000) was even more complex, involving other dimensions of dialogue function, including dialogue control, e.g., *greetings* and *introductions*, and task management, e.g., *deferring* and *closing*, but still with a dominant focus on illocutionary forces, e.g., *inform*, *commit*, and *suggest*. It is not difficult to see how this work relates to the problem tackled in this thesis. The names for these categories have a lot in common with verbs that label activity types in meeting summaries.

Early work on dialogue coding coalesced around the influential Discourse Resource Initiative program, which ran a series of workshops dedicated to exchanging information on discourse annotation schemes. This included schemes relating to illocutionary acts, discourse structure, coreference, and discourse segmentation. Part of the program involved proposing a standard scheme which eventually became DAMSL (Dialog Act Markup in Several Layers) (Core and Allen, 1997). This scheme adopted part of the multi-dimensional approach seen in conversation act theory (Traum and Hinkelman, 1992), calling the illocutionary components 'backward-looking' and 'forward-looking' functions, and employing tags dedicated to communication management as well. Since this influential work, a number of schemes have since emerged which have either taken this multi-dimensional tack, or collapsed the dimensions to produce a single layer. Prominent examples include COCONUT (Di Eugenio et al., 1997), SWBD-DAMSL (Jurafsky et al., 1997), DATE (Walker and Passonneau, 2001), MRDA (Shriberg et al., 2004; Dhillon et al., 2004), MALTUS (Clark and Popescu-Belis, 2004), and the AMI corpus scheme (AMI Consortium, 2008a).

Core problems with taxonomic dialogue coding

In the literature on dialogue coding, a set of core issues have continued to emerge which can be used to guide our approach. One of these is the problem of coding *reliability* (Carletta et al., 1997; Krippendorff, 2004; Artstein and Poesio, 2008). To obtain useful data, it is necessary for multiple annotators to agree on the assignment of categorical labels. But in the application of intentional or functional categories, annotators must rely upon mostly implicit cues from context. This makes such schemes especially difficult to apply reliably, particularly when a scheme has numerous or complex categories (Popescu-Belis, 2008).

It is in this relationship to complexity that reliability has a direct relationship to another problem facing dialogue coding—*coverage*. The development of taxonomies of intentional categories has a long history stemming from Searle's (1976) taxonomy of illocutionary acts. But despite extensive empirical work on the topic, there is still no general consensus on what constitutes a complete set of speech acts (let alone a complete set of communicative activities).[1] This is partly a result of a necessary tradeoff between reliability and coverage. Consider two prominent dialogue act schemes that have been applied to meetings. In the MRDA scheme (Shriberg et al., 2004), the set of categories was greatly expanded to assure coverage. In the AMI scheme (AMI Consortium, 2008a), the set of categories was greatly reduced to assure reliability. Neither of these claim to have achieved the ultimate goal—to be perfectly reliable and comprehensive. Instead, they take practical considerations more seriously than theoretical issues.

The coverage problem also relates to *multi-dimensionality*. In addition to requiring a wide variety of functions that can be applied reliably, there is general agreement that di-

[1] Whether such an endeavor is even appropriate is also an open issue.

alogue coding schemes should also accommodate multiple purposes for each utterance. This has been the focus of recent development of the DIT++ scheme (Bunt, 2008) and a related ISO standard (Bunt et al., 2010). To tackle this issue, the DIT++ scheme proposes a highly detailed division between dimensions, with a systematic relationship defined between them. This can make coding a laborious endeavor, requiring experts to perform the task.

Another issue that arises is the problem of appropriate *granularity* in discourse segmentation. Most existing coding schemes tend to operate at the level of individual utterances. This challenges their relevance to activity analysis, since many of the activities described in conversational summaries are summaries of extended sequences of interaction. Some dialogue coding has been done at larger granularities of analysis, but this has brought up some interesting issues of its own.

The scheme proposed by Ries (2001a) is a particularly interesting case—one that is perhaps the most relevant to the current thesis. First of all, the scheme is a dialogue coding that tackles segmentation and labeling problems at multiple levels, some of which have rather extended durations. The scheme proposes two structural levels above the utterance: a *topic segment* and a *document segment*. Each level of segmentation is associated with a taxonomy of labels. The document-level segmentation is associated with a “database” label which refers to what one might naturally call ‘genre,’ and includes categories such as *broadcast news* and *private telephone calls*. Document segments are also associated with “sub-database” labels that refer to sub-genres, e.g., *talk show* and *news show*. The topic-level segmentation refers to segments within individual conversations, and are assigned an *activity label* that includes categories such as *story-telling*, *discussing*, and *planning*. This latter level and its categories resemble the kind of dialogue activities evidenced in our meeting summaries.

Though it seems that the Ries (2001a) scheme might be useful for our current problem of analyzing meetings, one critical issue emerges that seems difficult to overcome—*domain dependence*. To exemplify, consider the fact that activities like *reviewing*, *seconding*, and *presenting* are important in our corpus of meeting minutes. Unfortunately, none of these are coded in the taxonomies of any previously designed coding scheme. Herein lies a critical problem with any taxonomic approach to intentional dialogue coding—there are many functions that one might assign to any instance of language use, and our natural understanding of each of those functions is vague and highly dependent on the situational context. Because of this, taxonomic approaches tend to fail when faced with new domains and new genres. Consider, for example, the 'introduction' activity in the AMI corpus discussed in Section 3.2. The participants' own summaries referred to this activity variously as "getting to know", "become acquainted with", and "introduce". All of these are 'correct' descriptions, but they call into question the nature of what distinguishes them. Is there a unifying notion underlying this activity? If so, is it a category of interaction that can be theoretically established? And would such a theoretical definition apply adequately to all arenas of language use?

These issues reflect an important underlying problem in operationalizing a computational model of communicative activities—the decision about which aspects of the model should be driven by theory and which should be driven by the data (or participants' own descriptions of it). For an activity-oriented meeting summarizer, for example, should one endeavor to design the activity types *a priori*? Or should one instead favor developing new categories as new observations are encountered?

An alternative to the taxonomic, computationally oriented approaches to discourse analysis may be found in the ethnomethodological approach taken by many sociologists and anthropologists. Research in this area suggests a more dynamic and flexible account than the taxonomic approaches mentioned above. The typically non-quantitative methods used, however, raise new issues about whether the approaches can be effectively computationalized. I shall cover this problem in the next section. For now, I summarize the contributions from this section with the following list, highlighting the difficulties that would face a taxonomic approach to activity coding.

Problem 1 It is difficult to simultaneously obtain both categorical nuance and coding *reliability* with taxonomic approaches.

Problem 2 It is difficult to establish full *coverage*, since the number of intentional categories can be large and discourse can be multi-functional.

Problem 3 It is difficult to determine the appropriate *granularity* for segmentation, since functional segments are widely varying.

Problem 4 It is difficult to apply taxonomic schemes to *new domains and genres*.

3.4.2 Activities as socially-constituted episodes of interaction

The categories employed by Ries (2001a) at the larger levels of discourse structure, such as *talk show* and *news show* suggest that as one ascends the discourse structure hierarchy, interpreting discourse purpose looks more and more like *genre* identification (Kessler et al., 1997; Wolters and Kirsten, 1999; Stamatatos et al., 2000; Dewdney et al., 2001). It is no surprise then, that communicative purpose is often used as the basis for *defining* genre (Swales, 1990; Kessler et al., 1997). But definitions of genre are notoriously hard to pin down. As Webber (2009) suggests, there is no one "right set" of features that can be used to define a genre. This is clearly true for meeting conversations. For example, how can one distinguish a 'debate' from a 'discussion?' And isn't it possible for an activity to be both a 'debate' *and* a 'discussion?' This problem becomes even more of an issue if the targets are natural descriptions of commonsense categories. It seems perfectly valid to say that there is no "right set" of communicative activity types, nor likely any "right set" of features for identifying them.

This suggests that it may not be appropriate to endeavor to taxonomize communicative activities at all. While the verbs in activity descriptions do establish a sort of activity typology, they do not establish a formal typology of the sort used in dialogue coding (or genre coding for that matter). It seems instead that they truly are natural labels, carrying with them all the vagueness and ambiguity that any natural language description entails.

It appears, then, that a study of the *features* of conversational activities is required before a model of activity types is proposed. Perhaps from a direct exploration of the patterns exhibited by features of the discourse, one can arrive at some schematic understanding of their nature and structure. Along these lines, there appears to be two possible approaches, both of which can provide an understanding of how people conceptualize types of activities. First, we can study the internal organization of activities, i.e., we can look at the observable *linguistic* and *interactional* properties of conversations, an approach that resembles research favored by corpus linguists doing genre analysis, e.g., Biber (1995). Second, we can study the *social* and *situational* contexts of conversation, an approach embodied by ethnographic analysis in sociolinguistics (Hymes, 1974).

I begin with an exploration of social factors. For this, I turn to Hasan (1991), who explores the factors that determine *how discourse types are constructed and distinguished*. In her analysis, Hasan does not develop specific categories of discourse types. Rather, she proposes a generic model for how discourse types emerge from what she calls a "generic structure potential." To build up the idea of a generic structure potential, Hasan begins with the notion of *cohesion*. In her model, discourse units are built from smaller units that are bound together by cohesion, i.e., relationships operating along different dimensions of discourse meaning. This idea is rather well known—it is the basis for the functional linguistic concept of 'lexical cohesion' (Halliday and Hasan, 1976).

For Hasan, lexical cohesion is a property that derives from something deeper. Instead of being grounded in sequences of topics and words, she proposes that discourse types are instead built from cohesion along the dimensions of "process, participants, attributes of participants, and the circumstances relevant to the process." For Hasan, this is called the *field* of discourse. It is not what people are talking about (i.e., topic), it is the people themselves, their relationships to each other, and the things that bring them together into coordinated action that create schematic discourse structures. In essence, Hasan's proposal is that discourse types derive from coherence relating to "our participation in everyday life in socially defined environments." It is this participation in a socially-constituted interaction that is the source of the semantic, referential, and lexical chains that can be observed.

Another line of work related to this is Korolija and Linell's (1996; 1998a; 1998b) studies of conversational *episodes*. Korolija (1998b) argues that spoken conversations are organized around *episodes*, which she defines as "gestalts of coherent dialogical interaction." She proposes a loose, multi-dimensional definition, that she contends is a necessary outcome of the complex socially-defined contexts of interaction. She observes that pauses, changes in topic, and other large-scale interactional properties in conversation are oriented principally around coherence in *participation frameworks*. Participation frameworks are configurations of participatory roles that maintain stability throughout an episode. They are constituted by *roles* relating to group action, as well as *attitudes* toward thematic content.

Hasan's idea of socially-defined types of "participation in everyday life" and Korolija's notion of episodic gestalts of interaction both echo a common theme in socio-linguistic analyses of communicative interaction—that communicative activities are built around a *commonsense, dominant social purpose*. Goffman (1979), for example, writes that we tend to perceive events in terms of what he calls "primary frameworks." In Goffman's model, primary frameworks are dominant configurations of role and purpose in social interaction. And he suggests they have a relationship to summaries as well, saying that "the type of primary framework provides a way of describing the event to which it is applied" (p. 24).

The primacy of higher-level social roles and relationships is also reflected in Clark (1996), who explains that we come together to communicate for "dominant purposes", and that these overarching goals are the main influence coordinating our actions. And Hanks (2005) shows that with each structural level of interaction comes a unique configuration of social roles and relationships, directly influencing the meanings of component utterances, and even individual words in use (e.g., *we*, *you*, *here*, and *there*).

Finally, we can turn to Levinson's (1992) seminal article on activity types. In his article, Levinson endeavors to give rigor to Wittgenstein's notion of "language games" by showing the substantial effect that activity type has on the contributions and inferences one can make in language use. In a lecture, for example, the instructor has prominent goals that are the overall purpose for the dialogue, such as providing an organized understanding of a body of knowledge. These goals then influence structures at smaller scales such as question-answer sequences, constraining their appropriate use and giving them special purpose within the larger interchange. A question-answer sequence within an airline ticket booking dialogue, for example, might exhibit a very similar local structure to one within a lecture, but the exchange will serve a totally different purpose and allow very different inferences to be drawn.

Two important conclusions can be taken away from the socially-oriented analysis of Levinson and the others just cited. The first is the notion that understanding speech acts and activities requires an account of their embedding within the context of social activity. This imbues every act of language use in a recursive, far-reaching, and ultimately *social* and *cultural* context by which its meaning is established. The second conclusion is derivative of the first—that the number and variety of activity types is limitless. Levinson proposes that Wittgenstein has it right in not making a distinction between speech acts and speech activities, suggesting there are countless varieties on countless scales, continually evolving along with our society and culture. Ultimately, Levinson (1992, p. 69) identifies activity types as "fuzzy categories whose focal members are goal-defined, socially constituted, bounded events with constraints on participants, setting, and on . . . the kinds of allowable contributions."

3.5 Outlining a technical approach: Meta-discourse in context

If the social, context-oriented perspective on discourse purposes suggests that activity categories are fuzzy and socially constituted (i.e., by participation and roles in interaction), how then does one proceed toward an observation-driven, *computable* operationalization? The system must somehow gain access to relevant social context, and it must also be accommodating to fuzzy categorizations. This must then be reconciled with the fact that summaries nonetheless have to make a choice about which activity-describing verb to use, thus realizing a form of hard categorization, albeit one that is based on the English lexicon. The verbs identified must somehow be used as representative of *commonsense* categories for activity types, and the central problem is inducing the meaning of such words in relation to how they are used to describe particular instances of activity. I shall now explore a methodological possibility for tackling this very problem, with the goal of narrowing in on a specific methodology to begin experimenting with.

The methodology I propose is based upon the observation-driven approach of *conversation analysis* (Sacks et al., 1974; Schegloff, 1980). Conversation analysis argues for analyses of structural patterns and social processes that are gleaned principally from evidence that the participants themselves have recognized them (Schegloff, 1993). The approach argues that because participants both create and deploy context in interaction, the only way to get a handle on an interactional phenomenon is to study how participants' behavior provides direct evidence for it. The idea is summarized well in the following passage:

> The best evidence that some practice of talk-in-interaction does, or can do, some claimed action, for example, is that some recipient on some occasion shows himself or herself to have so understood it, most commonly by so treating it in the ensuing moments of the interaction, and most commonly of all, next. Even if no quantitative evidence can be mustered for a linkage between that practice of talking and that resultant "effect," the treatment of the linkage as relevant—by the parties on that occasion, on which it was manifested—remains (Schegloff, 1993) (p., 101)

As applied to analyzing activity segments and activity types, the conversation analytic approach would therefore suggest that one must not look for theory-driven categorizations at all, no matter how fuzzy. Rather, analysis should employ participants' own speech as grounds for identification and characterization. Activity type labels can (and should) be derived from participants' own descriptions.

To apply this to our problem, we might look to instances when a participant might, for example, request that another participant "demonstrate" how something works. The ensuing dialogue then becomes direct evidence for what is meant by *demonstrate*, at least in that context. I refer to this kind of discourse about discourse as *bracketing meta-discourse* (Schiffrin, 1980), and this example might be called 'forward-looking'. But one can also imagine a 'backward-looking' form. A participant might summarize an interaction that just occurred by saying "Well, that was a healthy debate!" Using this kind of language provides direct access to participants' own appreciation of what is going on.

This thesis attempts to transform this simple idea into a practical quantitative method by finding statistical associations between meta-discourse and other observations in the surrounding dialogue. If those contextual observations are set up so they are relevant to identifying and characterizing activities, then meta-discourse becomes a reliable means for producing natural, observation-driven activity type labels. By virtue of associating common meta-discourse descriptions with learned groupings of feature distributions, this method could also allow identification of labels for specific occasions when participants *didn't* use meta-discourse.

What this approach embodies is an entirely observation-driven statistical analysis of activity segments, characteristics, and labels (i.e., an *unsupervised* machine learning methodology). A technical approach along these lines would have three basic components. The first concerns identifying the *occurrence* of an activity. By studying the distribution of social and participatory features in a dialogue (i.e., those features that correlate with activities), naturally occurring consistencies could be identified. If the features are chosen appropriately, one can hypothesize that these consistencies will allow identification of individual episodes of communicative activity. The second component relates to identifying *typical* patterns of such features across multiple instances of activity, i.e., *activity types*. Assuming that activity types do exist, one can hypothesize that there will be regularly identifiable groups of distributions of these features, and that these will correspond to the kinds of meaning people naturally use as a semantic representation for activities. Finally, the third component concerns meta-discourse, which could then be used to assign natural language labels to such types.

In the subsequent, final section of this chapter, I shall now explore which observable features are most likely to relate to the social and participatory meanings that underly communicative activities. These will then be drawn from in the experimental phase of research in order to formulate specific hypotheses about their usefulness as evidence for performing our three tasks.

3.6 Participant-Relational Analysis

In this section, I take the final step toward operationalizing the methodological suggestions described above. My aim is to propose a set of specific observable features that can be tested as indicators of the social and participatory characteristics of activities in conversation. These observable features, which I call *participant-relational features*, will form the basis of a concrete computational method for analyzing discourse segments, discovering activity types, and describing activities in subsequent chapters. I call the general approach presented here *participant-relational analysis*.

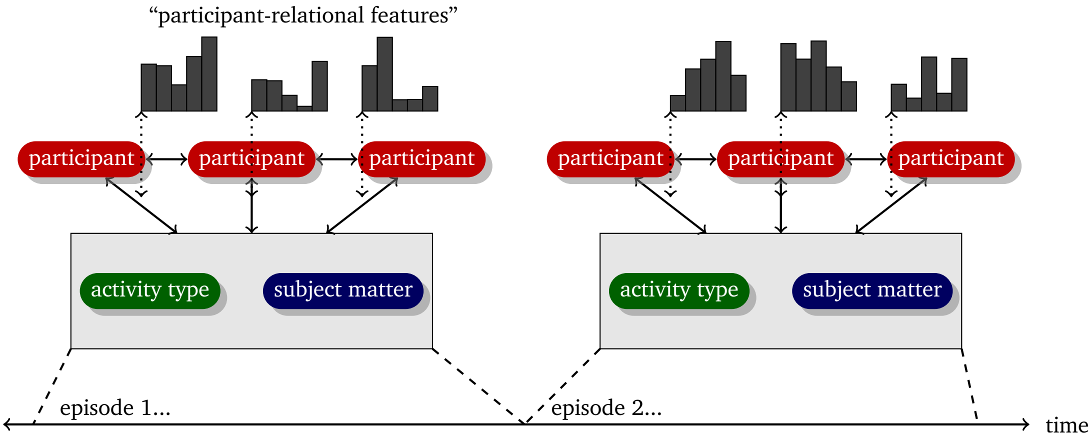
3.6.1 Introduction

Participant-relational analysis is grounded in two basic ideas. The first is that the activities we are interested in represent a coarse level of the *intentional structure* of dialogue (Grosz and Sidner, 1986). In other words, each activity is unified by a dominant purpose that is shared between the participants. It can then be inferred that there are linguistic properties reflecting this purpose which are shared amongst the utterances of a given activity episode.

The second idea concerns those specific properties that allow activities to be distinguished. Ethnographic approaches to discourse analysis suggest that activities center around participation in socially-defined contexts. Based upon this idea, it can be inferred that activities (and their types) may be distinguished according to two complex properties of utterances, both of which concern relationships between the participants and the utterance: *participant subjectivity* and *participant involvement*. Participant subjectivity concerns attitudinal, point-of-view, and other 'perspectival' relationships *toward* the dialogue content. This includes properties such as whether the utterance expresses the private mental state of the speaker, or the participants' temporal relationship to a described event. Participant involvement concerns the roles participants play *within* the dialogue content, e.g., as the agent of a described event.

Figure 3.6 sketches out the basic idea of participant-relational analysis. In essence,

Figure 3.6: A graphical representation of participant-relational analysis of activity episodes in conversation. Each activity is temporally associated with an episode of communication. Activities (and activity types) are identified and distinguished according to observable features indicating relationships between participants and the dialogue (and each other).



each activity in a conversation is associated with a unique configuration of participant relationships toward the dialogue (and the other participants). By identifying and counting features that associate with this configuration, one can distinguish activities and their types, and recover the components of meaning relevant to a summary for the activity.

It is important to highlight that this kind of analysis neglects consideration of relations between discourse segments, e.g, rhetorical or temporal relations between sentences, or the meanings of discourse connectives. These are clearly an important (if not central) factor in many theoretical models of discourse. The model just described instead makes a considerable simplifying assumption based upon a shallow *distributional* characterization of coherence within a segment. This is an important factor that limits the kinds of activities that might be explained by such a model. It is left for future work to integrate discourse segment relations and sequences into the proposed participant-relational method.

The work of Grosz and Sidner (1986) is a necessary starting point for understanding the approach. Their theory suggests that intentions (which equate to the goals and purposes of a dialogue) are a foundation for the structure of discourse. The natural aggregation of utterances into *discourse segments* arises from the fact that participants address a particular purpose for a period of time. The *attentional* state of the dialogue, which contains salient objects and relations, is then dependent upon the intentional and linguistic structure in the emerging dialogue, i.e., it is parasitic upon the underlying intentional structure. This notion has informed approaches to a variety of language processing problems, such as Centering theory (Grosz et al., 1995), which concerns the relationship of referring expressions to discourse coherence, and the idea that coreference and inferred relations between noun phrases are a useful basis for automatic intentional segmentation (Passonneau and Litman, 1997).

The participant-relational approach also employs this insight, but it highlights a unique observation—that objects in focus within the attentional state have an important quality which may be exploited—they are focused upon *by* the participants from particular *points of view*. In addition, the objects may in fact *be* the participants themselves. We would therefore expect the linguistic features which express such relationships to correlate with intentional structure, and to do so in a way which is important to participants' subjective experience of activity in the dialogue.

3.6.2 Generalizing the approach: Introducing the Pear Stories

To elucidate the idea of participant-relational features, I shall now introduce the Pear Stories corpus (Chafe, 1980b). Introducing this corpus here has two purposes. First, it provides an illuminating example of the relationship between participant-relational features and discourse structure. But secondarily, it also serves to demonstrate that the activity-oriented and participant-relational approaches are designed to apply not just to meetings, but to a wide range of both monological and conversational language. The proposed generality arises from being founded on the notion of a *joint* activity (Clark, 1996). Clark suggests that all language use, whether speaking to others, writing to an audience, or talking to ourselves, involves participants in joint action. In other words, there is always an addressee, and there is always a purpose for speaking. For this reason, it is suggested that activity-oriented summarization, including the participant-relational approach, ought to generalize to non-conversational domains.

We test this generalization using the Pear Stories for two main reasons. First, the Pear Stories contrast significantly with the AMI Corpus (described in depth earlier in this chapter). For example, the AMI Corpus is task-oriented whereas the Pear Stories are principally narrative. The AMI Corpus is multi-party whereas the Pear Stories are essentially monological. The second reason for using the Pear Stories is that there are a rich set of discourse segment annotations available. This make it a useful source for elucidating the theoretical contributions of this thesis, as well as for testing the theory experimentally.

The Pear Stories corpus consists of 20 spoken narrative monologues (Chafe, 1980b). To create the Pear Stories corpus, Chafe asked subjects to view a silent movie and then summarize it for a second person.[2] (Note that while only one person is speaking, there is another individual present to whom the speaker is providing the summary). The speech is manually transcribed and segmented into prosodic phrases. The corpus is also supplemented with a set of discourse segmentation annotations (Passonneau and Litman, 1997), where each monologue is segmented by seven annotators according to an informal notion of communicative intention. The segmentations embody an approach motivated by an *intentional* theory of dialogue structure (Grosz and Sidner, 1986), and the annotations were created by naive coders employing an informal notion of *speaker intention*.

The Pear Stories corpus has another interesting feature—that each monologue is itself a summary (of the Pear Stories movie). The Pear Stories subjects typically provide their summary in the form of a *narrative*, which one might consider to be the prototypical activity-oriented type of summary. However, our use of the Pear Stories is not for the purpose of studying narratives or narrative summarization. Rather, we approach the monologues themselves as objects to be summarized. As our analysis shows, it turns out that the Pear Stories participants are doing more than just "narrating" or "summarizing." Rather, they are performing many different types of distinguishable narrative-like activities, and we wish to study the mechanisms for distinguishing and summarizing them.

We now consider an example extract, shown in **(13)**, which highlights the relationship between participant-relational features and discourse structure. In the extract, two horizontal lines indicate a segment boundary which was identified by at least 3 of 7 annotators. A single horizontal line indicates a segment boundary which was identified by 2 or fewer annotators.

(11) [PEAR–09–21.2]
21.2 okay.

22.1 Meanwhile,
22.2 there are three little boys,
22.3 up on the road a little bit,
22.4 and they see this little accident.
23.1 And u-h they come over,

[2] The original Pear Stories movie, and Chafe's transcripts are publicly available on the Internet.

(11) [cont. . .]
23.2 and they help him,
23.3 and you know,
23.4 help him pick up the pears and everything.
24.1 A-nd the one thing that struck me about the- three little boys that were there,
24.2 is that one had ay uh I don't know what you call them,
24.3 but it's a paddle,
24.4 and a ball-,
24.5 is attached to the paddle,
24.6 and you know you bounce it?
25.1 And that sound was really prominent.
26.1 Well anyway,
26.2 so- u-m tsk all the pears are picked up,
26.3 and he's on his way again,

What is notable in this example is how changes in speaker intention correspond to changes in expressions of the speaker's relationship toward the content. In the example, there are three basic types of activity distinguishable according to the properties of participant subjectivity and involvement. The two segments beginning at 22.1 and 26.2 might be called a 'narrative' activity, and they share a similar use of the historical present tense (a type of participant subjectivity). Utterances 24.1 and 25.1, on the other hand, are a kind of 'reflection' activity, and unlike the other segments, they communicate *perceptions of the speaker*. The segment beginning at 24.2 could be called a 'generic description' activity, exhibiting its own distinct configuration of participant relational features, such as the *generic you* and *present tense*. It is also notable that participant-relational features relate directly to the very nature of the activity. In other words, the use of the historical present tense or perceptual language is more than just indicative of a 'narrative' or 'reflection' activity—it is *informative* of the character of those activity types (note the relation here to Borko and Bernier's (1975) typology of summarization systems). This suggests that participant-relational features can be used to characterize activities in addition to locating and distinguishing them.

3.6.3 Participant subjectivity and involvement

The term participant-relational, and the subcategories of participant subjectivity and involvement, are designed to be conceptual umbrella terms rather than formal theoretical objects. They originate from an attempt to unify and understand diverse accounts in which individual linguistic features have been shown to be indicators of discourse structures related to conversational activity (i.e., narrative segments, genre, interactional frameworks). One interesting example of this kind of work is Wortham (1996), who employs the distribution of personal pronouns and other deictics as coarse indicators of participants' conversational roles.

The theoretical contribution of participant-relational analysis may therefore be seen as a synthesis of disparate work along side an explanation of relevance to conversation summarization. In this section, I will survey this work from which the term is synthesized. I will discuss the specific features that have been found to correlate with discourse structure. The result will be an inventory of participant-relational features (one that is most certainly not exhaustive).

One particularly important source is Chafe (1994) and his work on the Pear Stories. He describes how speakers can express ideas from alternative perspectives, which he calls different types of "displaced consciousness." For example, a subject who is recounting the events in the Pear Stories movie has the option of saying "the man was picking pears", "the man picks some pears", or "you see a man picking pears." Each variant is an expression of the same idea but reflects a different perspective toward, or manner of participation in, the described event. He observes that discourse coheres in these perspectival terms, with shifts of perspective usually occurring at intentional boundaries. This particular example reveals specific linguistic variations, including *tense* and *aspect* variation in the main clause, and the use of a *generic pronoun* in a superordinate clause with a *cognitive verb*. The former variations are indicative of a speakers' temporal relationship toward a described event (Moens and Steedman, 1988). This is an example of what I call *participant subjectivity* because it reflects a subjective attitude or perspective toward the content. The latter directly encodes the speaker (albeit generically) into the description, and it is an example of what I call *participant involvement* because reference to the participant plays an explicit role in the construction.

Wiebe (1994; 1995) has investigated a phenomenon closely related to this—*point-of-view* in fictional narrative. She notes that paragraph-level blocks of text often share a common *objective* or *subjective* context. That is, sentences may or may not be conveyed from the point-of-view of individuals, e.g., the author or the characters within the narrative. Sentences continue, resume, or initiate such contexts, and she develops methods for automatically determining when the contexts shift and whose point-of-view is being taken. A survey of the features she employs shows that most relate to the participants' relationship to the content. For example, she employs *spatial deixis* and *temporal deixis* as indicators of psychological point of view. *Personal pronouns* play a role in detecting the experiencer of subjective expressions. And she proposes that subjective elements can include *evidentials* expressing certainty or uncertainty, *past perfective* and *progressive* aspect, and *kinship terms*, just to name a few.

Smith's (2003) analysis of texts draws a more general set of connections between the content of sentences and generic types of large-scale discourse segments. She does this by analyzing texts at the level of short passages and determines a non-exhaustive list of five basic "discourse modes" occurring at that level: narrative, description, report, information, and argument. These can be seen as a textual variant of the communicative activity types of interest in this study, with a clear relationship to genre analysis as well. In Smith's model, the discourse mode of a passage is determined by the type of situations described in the text (e.g., event, state, general stative, etc.) and the temporal progression of the situations in the discourse. Relevant to the participant-relational approach, Smith's situation types are organized according to the perspectival properties of aspect and temporal location. A narrative passage, for example, relates specific events and states, with dynamic temporal advancement of narrative time between sentences. On the other hand, an information passage relates general statives with atemporal progression. This provides us with motivation that activity episodes might contain consistent distributions of tense and aspect features.

The three studies just discussed tend toward a linguistically-driven analysis, mainly focused on specific types of written language. To understand the problem from a more sociologically and interactionally oriented viewpoint, it is helpful to consider Goffman's (1974) studies of experiential *framing*. In Goffman's terms, framing has to do with the fact that when participants interact, their actions take place within a universe of prior social experience, often glossed simply as *context*. Understanding this context is a necessary condition for interpreting any communication because it has a fundamental influence on interpretation on every level.

What Goffman brought to the table in the discussion of framing is that context is made evident through its direct invocation in the interaction itself. To exemplify, consider a person drawing a diagram on a whiteboard who is suddenly interrupted by another person who says “No, that won’t work. Let’s start over.” By observing this interaction (which contains no reference to subject matter at all), one suddenly learns a lot about what these people are doing (i.e., it’s unlikely to be a classroom lecture, and more likely to be a collaborative design meeting). This is indicated principally by the word *let’s* and reference to future events. Or consider one commonly encountered type of framing—*quoted speech*, e.g., “He said, ‘you better give that to me now!’” The basic meaning of quoted speech utterances must be gotten by complex mental transformation of who is participating and what they are doing. This is often explicitly cued by phrases like “He said.” The same applies in ever more subtle ways to story-telling, theatrical performance, and political speeches. All of our everyday conversations are embedded within such participatory transformations, and our interactions cue them directly.

To appreciate Goffman’s treatment of interaction and framing, and to understand which features are indicative of it, it is necessary to develop a richer meaning for the concept of *participation*. For Goffman, participation is more than the commonsense notion of ‘being involved.’ Rather, it is an analytic concept that stands for the relations that a participant has toward other participants and the activity, their roles, attitudes, and purposes. All participants come into activity with such relationships, and the established complex network of such relations constitutes what may be called the *participation framework*.

Participation frameworks are thus a mechanism whereby a division of labor is established for an activity. Interactionally speaking, participation frameworks are important because different roles require unique interactional responsibilities (Clark, 1992). One linguistic phenomenon that is generally relevant to this is *deixis* (Fillmore, 1997; Hanks, 2005). It is a common assumption in computational studies of dialogue (e.g., (Poesio, 2004a; Nissim et al., 2004)) that the word *you* refers either generically or to the addressee(s). A richer notion of participation frameworks, however, shows us that this is a problematic assumption. Instead, participants are able to project themselves as social and moral actors using a diverse array of epistemic, moral and affective stances (Goodwin, 2007), and this influences even the most basic aspects of linguistic interpretation. The words *you* and *we*, for example, can be used for generic, plural, or singular reference, as addressee-inclusive or addressee-exclusive, in reference to hypothetical individuals or non-human entities, or even metonymically in reference to objects connected to individuals (Mühlhäusler and Harré, 1990; Wales, 1996). This suggests that *person reference* is perhaps the most important type of participant-relational feature, as it is the principal means for instantiating participants into interactional roles.

It should be clear from this discussion that participant-relational features are a diverse collection, unified by a rather generic notion—that anything that indicates a participant's attitude, perspective, point-of-view, or participation in the dialogue content is fair game. In other words, any encoded relationship is useful. Grammatically speaking, this tends to emerge in a variety of ways, including strictly deictic expressions, either temporal (*now* and *then*), spatial (*here* and *there*), or personal (*we* and *you*). It emerges in less strict notions of deixis (Fillmore, 1997), such as tense and aspect. It also emerges in phrases that frame sentences explicitly, such as hedges (e.g., *maybe*), epistemics (e.g., *I think*), and markers of quoted speech (e.g., *he said*). One might also extend this idea to ever more subtle markers of participant relationships, such as grammatical *mood* and evaluative expressions (e.g., *good* and *terrible*).

The notion of a participant-relational feature is intended as basic guidance in the modeling of communicative activities. The intention is to take a simple step forward by synthesizing results in linguistics, anthropology, psychology, and sociology in such a way that they can be applied simply in a computational, statistical setting. In the following chapters, I select a collection of participant-relational features and investigate their practical use within a quantitative, experimental setting. It is there that I present specific hypotheses and describe the manner in which they will be tested.

3.7 Summary

In this chapter, I have described the essential properties of an *activity-oriented* analysis of meeting conversations (i.e., a summary that focuses on what 'happened' or what the participants 'did,' rather than what the conversation was 'about'). I have shown that meeting summaries commonly refer to the activities of participants, particularly communicative activities that occur as conversational episodes. I have proposed a prototypical semantic structure for activity descriptions that includes three parts: an activity type, a configuration of participatory roles, and subject matter. By turning to ethnomethodological, social constructionist analyses of discourse function (in contrast to computational, taxonomic approaches), I suggested that activities occur as socially-constituted categories with fuzzy, dominant social purposes. I then sketched out a methodological framework for how one might analyze those activities in a way that leverages participants' own expression to arrive at an appropriate description. In the final section, I proposed an abstract category of linguistic indicators called participant-relational features, which serves as a conceptual guide in the selection of specific concrete features to test.

It is at this point that I now turn to the quantitative phase of this thesis. The following chapters contain a collection of experiments in automatic analysis of conversations that test and validate the participant-relational approach.

Chapter 4

Annotation and Evaluation: Pre-requisites to the Main Experiments

In this chapter, I present two important validations of the experimental methods used in the thesis. The first of these helps to validate the use of participant-relational features, which are to be used as input to the summarization algorithms being tested. Specifically, I present an annotation and coding reliability study (Section 4.1) of verbal reference to participants. The annotation scheme introduces novel distinctions for vagueness and addressing-based referent analysis, and its application to a corpus of meetings provides a deeper understanding of participation in dialogue and confirms reliable coding of participant reference in workplace meetings. The second study (Section 4.2) assesses the appropriateness of commonly-used evaluation measures in discourse segmentation (the principal task studied in the experiments). The study shows that the discourse segmentation performance measures $\mathrm{P_k}$ *(Beeferman et al., 1999) and* WindowDiff *(Pevzner and Hearst, 2002) are biased in favor of segmentations with fewer or adjacent segment boundaries. This results in their failure to penalize substantially defective segmentations. To resolve this problem, I propose a novel unbiased measure* k-κ, *which accounts for chance agreement. I go on to replicate a recent topic segmentation experiment (Eisenstein and Barzilay, 2008), drawing substantially different conclusions about the effectiveness of state-of-the-art segmentation algorithms. These results help to determine which algorithms are best suited for application in subsequent participant-relational segmentation experiments.*

In the previous chapter, a collection of novel codings of linguistic content were proposed. Namely the general concept of a participant-relational feature was described, and an argument was given to support its use in performing activity-oriented summarization. Because of the novelty of participant-relational feature coding, and the new *annotation* tasks that will arise from its use, it is important to validate that such codings can be performed reliably before experimenting with them as a means for summarization. In Section 4.1 of this chapter, a case study in participant-relational coding is provided to this end. It presents a novel method for annotating references to people in conversation, with a focus on references to the *participants* themselves. Performing this study is important because all participant-relational features ultimately rely on understanding how participants are indexed in elements of language. Confirming that reliable annotation of participant reference is possible helps to establish a more solid footing for the future design of other participant-relational coding schemes.

Another pre-requisite to the experiments in this thesis, which focus principally on discourse segmentation, is that of determining the appropriate form of segmentation evaluation. Specifically, one must confirm that the chosen measure of performance evaluates the appropriateness of the segmentation in an unbiased way. Section 4.2 in this chapter presents such a confirmation in the form of a novel mathematical analysis of commonly-used *segmentation evaluation methods*. The study shows that these commonly-used methods are in fact substantially flawed due to biases toward certain types of segmentation. State-of-the-art segmentation algorithms are therefore re-evaluated and a new segmentation evaluation measure is proposed.

Together, the two studies in this chapter help to validate the participant-relational approach and give us confidence that our experimental methods are valid.

4.1 Participant-Relational Annotation: Case Study in Participant Reference[1]

To realize an activity-oriented, participant-relational approach to summarization, understanding reference to participants is of central importance. The approach requires understanding roles and relationships in a dialogue, and it must interpret their linguistic expression in the dialogue. An activity-oriented summarizer therefore requires, at the least, understanding references to participants and recognizing discourse structure through evidence of participant-referential coherence.

With this requirement in mind, this section describes research on designing and applying a *person reference* annotation procedure. The procedure is a coreference annotation scheme that focuses on distinctions between different types of *participant reference*, the

[1] A condensed version of this section was published as “Annotating Participant Reference in English Spoken Conversation” in the Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pp. 256–264, Uppsala, Sweden, 15–16 July 2010.

predominant type of person reference in face-to-face multi-party conversation. (The complete annotation guidelines, and a reference sheet for those guidelines are presented as Appendices A and B).

Participant reference is exemplified by the use of proper names such as *James*, or most commonly by the pronouns *I*, *you*, and *we*. It plays an essential role in many of the most important types of expressed meanings and actions in conversation, including subjective language, inter-personal agreements, commitments, narrative story-telling, establishing social relationships, and meta-discourse. In fact, some person-referring words are the most frequent words in conversation. The words *I* and *you* are the most frequently used nominals in several conversational corpora, including Switchboard (Godfrey et al., 1992) and the AMI Meeting Corpus (Carletta, 2007). In the British National Corpus they are the two most common of any words in the demographic (i.e., conversational) subcorpus (Burnard, 2007). And Google's Web 1T 5-gram statistics (Brants and Franz, 2006) list *I* and *you* as more frequent even than the word *it*. The word *we* falls within the top 10 most frequent words in all of these corpora.

Perhaps contrary to intuition, however, interpreting person-referring expressions can be rather complex. Person-reference interpretation is strongly dependent on social, situational, and discourse context. The words *you* and *we* are especially problematic. Either can be used for generic, plural, or singular reference, as addressee-inclusive or addressee-exclusive, in reference to hypothetical individuals or non-human entities, or even metonymically in reference to objects connected to individuals (Mühlhäusler and Harré, 1990; Wales, 1996). In addition, these and many other issues are not simply occasional problems but arise regularly.

A major aim of this study is therefore to increase understanding of the nature of participant reference through presenting a nuanced yet reliable set of type and property distinctions. The scheme proposes novel distinctions concerning three main factors. The first distinction concerns vagueness and indeterminacy, which is often exploited by speakers when using words such as *you*, *they*, and *we*. The scheme aims to provide a reliable basis for making an explicit distinction between specific and vague uses. The second distinction concerns an issue faced frequently in informal conversation, where words typically used to do person-referring are also commonly used in non-person-referring ways. Another aim is therefore to establish reliable person/non-person and referential/non-referential distinctions for these words. The third issue concerns what I call 'addressing-based inclusion attributes'. Qualitative analysis suggests that participant frameworks (participant roles in the ongoing conversational activity) are often expressed using vague or underspecified references to participants, where the speaker, addressee, or others are implied (often subtlely) as members of the group of referents. Understanding these local expressions of participatory context may therefore provide insight into the organization of communicative activities. The scheme therefore goes beyond the specific/underspecified/generic distinctions made in schemes such as ACE (NIST, 2008), in order to study participant membership in person referring expressions and its relationship to participation structure.

To exemplify some of these issues, consider extract **(12)** from the AMI corpus, which based upon first-hand experience with the corpus is typical in terms of complexity of person-reference.

(12) [AMI–ES2004b–249]
249 D: Um, current remote controls do not match well with the operating behaviour of the **user** overall.
250 D: For example, **you** can see below there, seventy five percent of **users** zap a lot,
251 D: so **you**'ve got **your person** sunk back in the sofa channel-hopping.

As this example demonstrates, person-referring expressions have many potential meanings and are often vague or non-specific. In this case, "the user" refers to a non-specific representative of a hypothetical group. The group itself is then referred to directly as "users." The first use of "you" refers to the addressees, but the second 'generic' use has a speaker and addressee-inclusive meaning (though it can be argued that it does not refer to all individuals generically). The phrase "your person" refers to a specific hypothetical example of the "users" referred to previously.

Clearly, interpreting the person references in this extract depends upon a comprehension of the situational and social context of the interaction. Some instances are also likely to have no singular 'true' meaning. Rather, different individuals will infer slightly different referential meanings. The goal of the scheme presented here is to study this rich, context-driven interpretation problem and to use the result as a means of access to participant's subjective view of their participation in the communicative activity.

4.1.1 Background and related annotation schemes

Previous work on reference annotation has covered a wide range of issues surrounding reference generally. It is useful to categorize this work according to the natural language processing tasks the annotations are designed to support.

Schemes for discourse analysis

Several reference annotation schemes have been designed with the goal of testing linguistic theoretical models of discourse structure or for use in the study of discourse processing problems like anaphora resolution and reference generation. These schemes have been applied to both text and dialogue and label discourse references with a rich set of syntactic, semantic, and pragmatic properties. For example, the DRAMA scheme (Passonneau, 1997) and the GNOME scheme (Poesio, 2000, 2004a) include labels for features such as bridging relation type and noun phrase type in addition to a rich representation of referent semantics. Other schemes label animacy, prosody, and information structure to study their relationship to the organization and salience of discourse reference (Nissim et al., 2004; Calhoun et al., 2005). Recent developments include the explicit handling of anaphoric ambiguity and discourse deixis (Poesio and Artstein, 2008).

Despite the depth and detail of these schemes, participant reference has not been their main concern. The annotations by Poesio (2000, 2004a) include dialogue source material, but the rather constrained interactional situations do not elicit a rich set of references to participants. The scheme thus employs simple default labels for words like *I* and *you*. The work by Nissim et al. (2004) is an annotation of the Switchboard corpus (Godfrey et al., 1992), which contains only two participants who are neither co-present nor socially connected. Participant reference is thus rather constrained. Other than labeling coreferentiality, the Nissim scheme includes only a single distinction between referential and generic instances of the word *you*.

The only other annotations involving participant reference of which the author is aware are a set of addressee annotations (Jovanovic et al., 2006; Jovanovic, 2007) of the AMI corpus, which despite being quite large and involving our corpus of interest, do not provide precisely the type of annotations we are interested in here. Rather, these annotations label each utterance with its intended addressee(s).

Schemes for information extraction

In contrast to the schemes described above, which are mainly driven toward investigating linguistic theories of discourse processing, some reference annotation projects are motivated instead by information extraction applications. For these projects, a priority is placed on entity semantics and coreference to known entities in the world. For example, the objective of the Automatic Content Extraction (ACE) program (Doddington et al., 2004) is to recognize and extract entities, events, and relations between them, directly from written and spoken sources (mostly from broadcast news). The schemes thus focus on identifying and labeling the properties of entities in the real world, and then marking expressions as referring to these entities. Recent work in the ACE project has expanded the scope of this task to include cross-document recognition and resolution (Strassel et al., 2008). In the ACE scheme (NIST, 2008), person reference is a central component, and in the broadcast conversation component of the corpus there is an extensive inventory of participant references. The annotation scheme contains a distinction between specific, underspecified, and general entities, as well as a distinction between persons and organizations.

Another closely related set of studies are four recent investigations of second-person reference resolution (Gupta et al., 2007a,b; Frampton et al., 2009a; Purver et al., 2009). These studies are based upon a common set of annotations of the word *you* in source material from the Switchboard and ICSI Meeting corpora. The purpose for the annotations is to support learning of classifiers for two main problems: disambiguation of the generic/referential distinction, and reference resolution for referential cases. In addition to the generic/referential distinction and an addressing-based reference annotation, the scheme uses special classes for reported speech and fillers. It also allows annotators to indicate vague or difficult cases. The current work builds directly upon this prior work by extending the annotation scheme to *all* person-referring expressions.

4.1.2 Annotation method

The current study follows an annotation method consisting of two main phases: a preliminary phase where the first names of the conversation participants are identified, and a subsequent person reference annotation process. The first phase is not of central concern to the thesis, though a brief summary is provided below. The primary focus is the second phase, during which every instance of person-referring occurring in a given conversation is labelled with several properties. The discussion in this section shall focus on the novel and challenging aspects of this multi-step annotation process. A brief summary of the annotation tool is also presented.

Source material

The annotation scheme was applied to source material drawn from two source corpora: the AMI corpus (Carletta, 2007), which contains experimentally-controlled scenario-driven design meetings, and the ICSI corpus (Janin et al., 2003), which contains naturally occurring workplace meetings. All the meetings have at least four participants and have an average duration of about 45 minutes. In the AMI corpus, the participants are experimental subjects who are assigned institutional roles, e.g., project manager and industrial designer. This helps to establish controlled social relationships within the group, but generally limits the types of person referring. The ICSI meetings are naturally occurring and exhibit complex pre-existing social relationships between the participants. Person referring in this corpus is quite complex and often includes other individuals from the larger institution and beyond.

Phase 1: Labeling participant names

The first phase of annotation consists of identifying the names of the participants. This task is performed for every participant in every meeting in the AMI and ICSI corpora, which totals 275 unique participants in 246 meetings. Despite the fact that the participants' are given anonymized identifiers by the corpus creators, determining participants' names is possible because name mentions are not excised from the speech transcript. This allows identification of the names of any participants who are referred to by name in the dialogue, as long as the referent is disambiguated by contextual clues such as addressing.

To extract name information, the list of capitalized words in the speech transcript was scanned manually for likely person names, generating a list of *probable person names*. This was done manually due to the difficulty of training a sufficiently robust named-entity recognizer for these corpora. The process then involved iterating through three main steps for each unnamed person: (1) identifying (using the anonymized person identifiers) any meetings in which the unnamed person participated, (2) considering each probable person name in order of frequency of occurrence in those meetings, and (3) replaying short snippets of the recording surrounding occurrences of that token until a name could be matched to the anonymized identifier.

In most cases, a name was used in reference to a participant at some point, and when that instance was found, the discourse context made it clear which participant was the intended referent. In the AMI meetings, 158 of 223 (71%) of the participants' first names were identified. In the ICSI meetings, 36 of 52 (69%) were identified. While these numbers may seem low, failure to determine a name was generally associated with a low level of participation of the individual either in terms of amount of speech or number of meetings attended. As such, the proportion of utterances across both corpora for which the speaker's name is identified was actually 91%.[2]

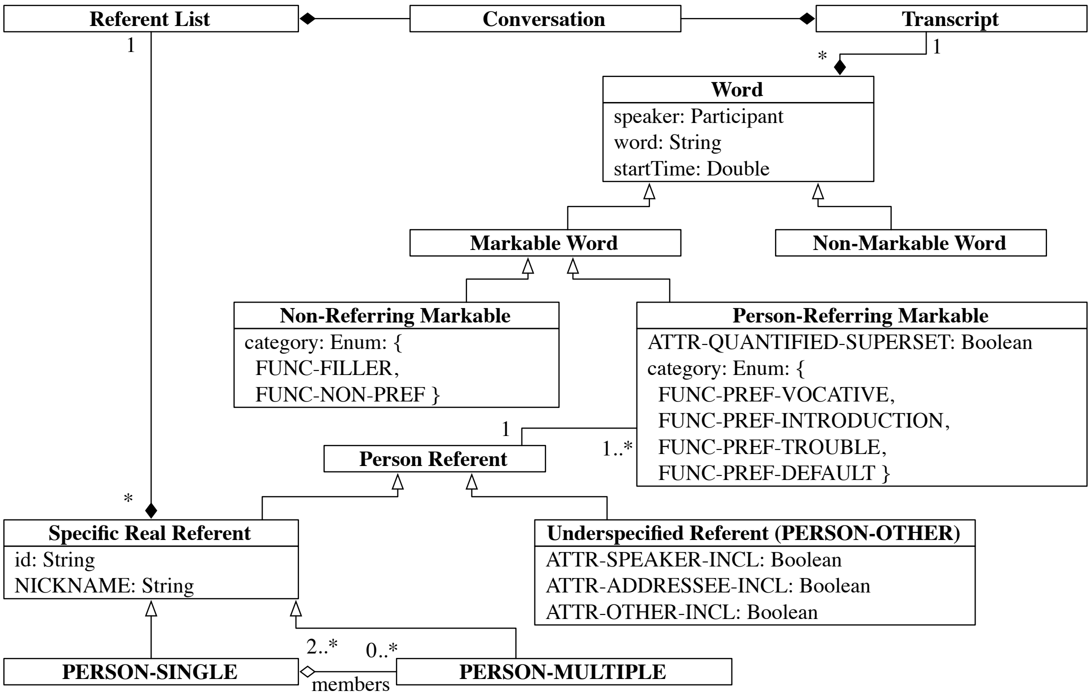
Phase 2: Person-reference annotation

The second phase of annotation consisted of annotating **person-referring**, i.e., instances of verbal reference to people. In practice, the procedure was divided into four main annotation steps: markable identification, referent identification, functional category labeling, and coreference linking. For non-specific references, the last step also involved labeling addressing properties. For each annotated conversation, annotators labeled every instance of person-referring in every utterance, performing the steps in chronological sequence for each utterance. The Unified Modeling Language (UML) diagram in Figure 4.1 depicts the formal data structure produced by the procedure. The diagram may also be viewed as approximating a decision tree for the main annotation steps.

Step 1. Markables. The first step was markable identification, which involved recognizing person-referring expressions in the transcript. (This is represented by the distinction between the *Markable Word* class and the *Non-Markable Word* class in the UML diagram.) Only expressions that are noun phrases were considered, and only the head noun was actually labeled by the annotator (the extent of the expression was not labeled). These identified head nouns are called **markables**. Note, however, that before human annotation began, an automatic process was used to identify likely occurrences of person-referring expressions. The process identified all personal pronouns as markables, except *it*, *them*, and *they* (since these are more likely to be non-person-referring) and the *wh*-pronouns (not labeled in the scheme). It also marked any occurrences of the previously identified partici-

[2] Though not presented in this thesis, a preliminary experiment in person name induction has been conducted using these data. The goal of these experiments is to replicate this human labeling process automatically, using knowledge about addressing behaviour and the use of vocatives and introductions (see below for some discussion of these labels).

Figure 4.1: A Unified Modeling Language (UML) diagram depicting the formal structure of the produced annotation data. The structure is organized around two main elements: the transcript (containing words) and the referent list (containing specific real referents). Annotation decisions made by annotators assign elements of the text (and referents referred to in the text) to classes in this structure. This establishes a correspondence between steps in the annotation process and components of the diagram, similar to a decision tree. These correspondences are described in Section 4.1.2. In the diagram, filled diamonds indicate composition between objects; open diamonds indicate aggregation between objects; no decoration indicates a generic association between objects; open triangles indicate a sub-class relation between two classes.



pants' proper names. Of course, some of these automatically identified words might *not* be person-referring, or there may be instances of person-referring that are *not* automatically identified. Annotators did not unmark any of the incorrect positively identified words, even if they are not person-referring. Rather, in the subsequent step, they were labelled explicitly as person-referring or not. The resulting set of manually and automatically identified words constituted the complete set of markables.

Step 2. Person Referents. The second step was the identification and labeling of person referents. (This is represented by the *Referent List* collection in the UML diagram, which is composed of *Specific Real Referent* objects.) Any people or groups of people that were referred to specifically and unambiguously (see below for definitions) were added by the annotator to a conversation **referent list**. The list was automatically populated with each of the conversation participants, and annotators were also able to assign nicknames for these referents.

Step 3. Reference Function. The third step consisted of labeling markables with a **functional category** (FUNC-*). (This is represented in the UML diagram in the distinction between *Non-Referring* and *Person-Referring* markables, as well as in the enumerated subcategories within those two boxes.) The functional categories serve two main purposes. First, they are used to distinguish person-referring markables from all others. Second, they are used to distinguish between specific dialogue purposes, such as whether the reference is used in an introduction, a vocative, or a filler.

Step 4. Coreference and Addressing Inclusion Attributes. The final step was to link the markables that were labeled as person-referring to the appropriate referent in the referent list. (This is represented as the association between *Person Referent* and *Person-Referring Markable* in the UML diagram.) This is only done for specific and unambiguous referring. Otherwise, the referent is said to be underspecified, and instead of linking the markable to a referent, it is labeled with three binary addressing inclusion attributes (see the properties in the *Underspecified Referent* box in the diagram). **Inclusion attributes** label whether the speaker, addressee, or any other individuals are included in the set of people being referred to, given the social, situational, and discourse context (details below).

Detailing the important distinctions

Defining 'person' and 'referring'. To be **person-referring**, an expression must satisfy two conditions. First, the expression's primary contribution to the *speaker's intended meaning or purpose* must be either to identify, label, describe, specify, or address. These are the basic types of **referring**. Second, the referent being identified, labeled, etc., must be a **person**, which is defined to include any of the following: a distinct person in the real world; a fictitious or hypothetical person; a human agent, perceiver, or participant in a described event, scene, or fact; a class, type, or kind of person, or representative thereof; a specification or description of a person or set of people; a (possibly vaguely defined) group or collection of any of the above; the human race as a whole, or a representative thereof.

If a noun phrase is used to do person-referring as defined, the associated markable is labeled with one of the four person-referring functional categories (FUNC-PREF-*). If a markable is not person-referring (either non-referring or referring to a non-person referent), it is labeled with the functional category FUNC-NON-PREF. The one exception to this is the use of a pre-defined list of common discourse fillers such as *you know* and *I mean*. When used as fillers, these are labeled with the non-referential FUNC-FILLER category.

Joint action and referring 'trouble'. Annotators are asked to consider occasions of referring to be *joint* actions between the speaker and the addressee(s) of the utterance. The annotator assumes the role of an overhearer and considers as referring any case where the speaker's *intended purpose* is to refer. If the instance of referring is not successfully negotiated between the participants (i.e., common ground is not achieved), but the speaker's intended purpose is to refer, then the annotator marks this as FUNC-PREF-TROUBLE. This is used to identify problematic cases for future study.

Specific, unambiguous referring. Only the referents of *specific, unambiguous referring to a person in the real world* (PERSON-SINGLE) are included in the conversation referent list and made the subject of coreference annotation. References to more than one such individual can qualify (PERSON-MULTIPLE), but only if the members are *precisely enumerable* and qualify individually. The motivation for this distinction is to distinguish references that would be directly useful to applications. Coreference for underspecified references is not labeled.

Special functional categories. Two functional categories are used to distinguish special uses of person-referring for possible future use in speaker name induction (the task of automatically learning participants' names). The two categories are FUNC-PREF-INTRODUCTION and FUNC-PREF-VOCATIVE, which specify personal introductions such as "Hi, I'm *John*," and vocative addressing such as "What do you think, *Jane*?" These categories are used only for proper names.

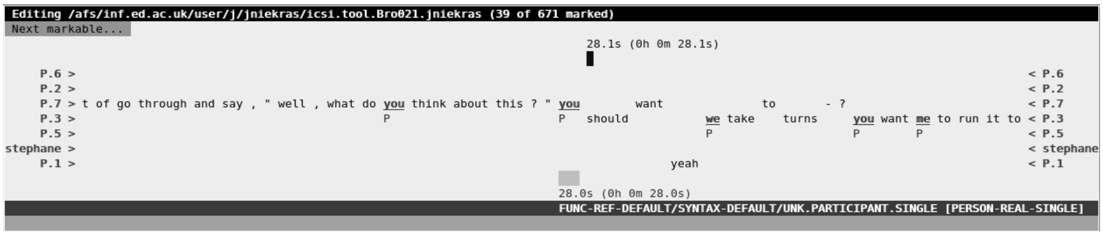
Addressing-based inclusion attributes. A major novelty in the scheme is the use of addressing-based distinctions for underspecified referents. Rather than using the labels 'generic' or 'indeterminate', three binary attributes are employed (ATTR-*-INCL) that label whether the speaker, addressee or any other real individuals are members of the set of people referred to.

The use of this distinction is informed by the notion that addressing distinctions are of central importance to the recognition of activity type, structure, and participation roles. A generic pronoun, for example, will often have all three categories labeled positively. But as an example of where this scheme creates a novel distinction, consider the phrase "You really take a beating out there on the pitch!", where the speaker is a football player describing the nature of play to someone who has never played the game. This 'generic' use of *you*, used in an activity of autobiographical description, is intuitively interpreted as not including the addressee (ATTR-ADDRESSEE-INCL=FALSE) but including the speaker and others (ATTR-{SPEAKER,OTHER}-INCL=TRUE). These distinctions are hard to motivate grammatically yet critical to identifying useful properties relating to participation in the communicative activity.

Unknown specific referents. In some cases, an annotator can determine that a reference is specific and unambiguous for the participants but the annotator himself is unable to determine the identity of the referent. This is generally due to a lack of contextual awareness such as not having adequate visual cues. In such cases, the annotator assigns a special REF-UNKNOWN referent.

Quantified expressions. Quantified expressions are another case that is handled in an unusual way. Annotators are asked to interpret the entire NP and determine whether the referent of the entire phrase is specific and unambiguous. If so, the phrase is annotated as

Figure 4.2: A screenshot of the person reference annotation tool. Each participant's speech is displayed on its own line, and a cursor follows along in sync with the audio. Annotation labels are dynamically revealed in the bottom rows as the user scrolls.



```
 Editing /afs/inf.ed.ac.uk/user/j/jniekras/icsi.tool.Bro021.jniekras (39 of 671 marked)
 Next markable...
                                                                    28.1s (0h 0m 28.1s)
                                                                    █

    P.6 >                                                                                                       < P.6
    P.2 >                                                                                                       < P.2
    P.7 > t of go through and say , " well , what do you think about this ? " you       want          to      - ?   < P.7
    P.3 >                                         P                        P  should           we take   turns    you want me to run it to < P.3
    P.5 >                                                                                  P                 P       P  < P.5
stephane >                                                                                                      < stephane
    P.1 >                                                                        yeah                           < P.1

                                                                   28.0s (0h 0m 28.0s)
                                                                   FUNC-REF-DEFAULT/SYNTAX-DEFAULT/UNK.PARTICIPANT.SINGLE [PERSON-REAL-SINGLE]
```

a whole and quantification is subsumed into the interpretation. Otherwise, the annotator annotates the phrase according to the set of individuals over which the quantifier is scoped and the attribute ATTR-QUANTIFIED-SUPERSET is assigned.

Other difficult aspects of the annotation procedure are covered in the annotation manual, including handling of disfluencies and identifying lexical heads.

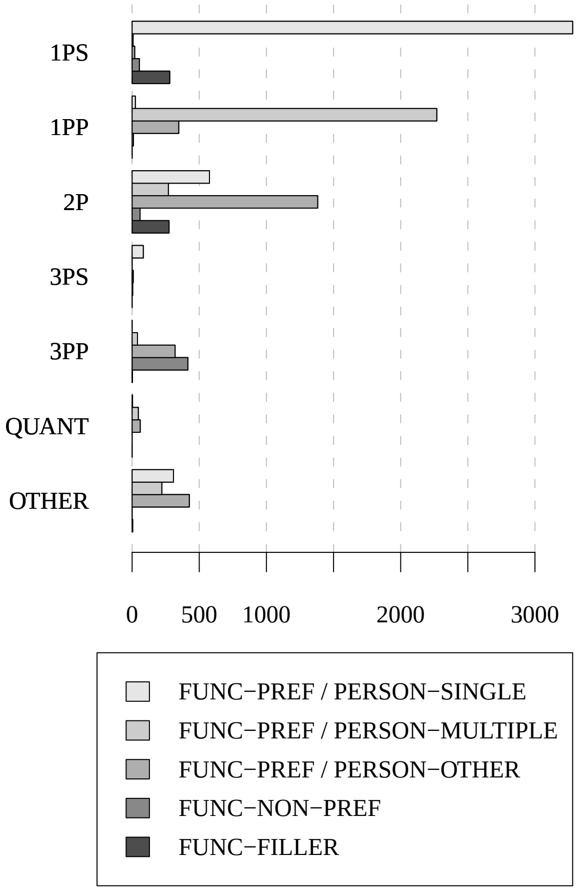
Annotation tool

The annotations were collected using a purpose-built software tool designed for discrete event-based annotation of multi-modal corpora. The tool uses a simple, low-latency text-based interface that displays multiple streams of discrete events in temporal order across the screen. In this case, the events are time-synchronized words that are distributed to different rows according to speaker. The interface allows keyboard input only and is synchronized with the MPlayer playback engine. A screenshot of the interface is shown in Figure 4.2.

4.1.3 Annotation results and analysis

The dataset produced by the study consists of approximately 11,000 individually annotated referring expressions in 16 meetings from the AMI corpus (Carletta, 2007) and 3 meetings from the ICSI corpus (Janin et al., 2003). Figure 4.3 shows, for each grammatical type of referring expression, the frequency of five principal markable types, organized as follows. First, the two non-person-referring functional categories FUNC-NON-PREF and FUNC-FILLER are included as annotated. Second, a three-way breakdown of person-referring according to the type of person referent is included: a specific individual (PERSON-SINGLE), multiple specific individuals (PERSON-MULTIPLE), or underspecified (PERSON-OTHER). The grammatical types include a grouping of the personal pronouns by grammatical person and number

Figure 4.3: The frequency of referring types in the person reference corpus, by grammatical type of the referring expression. Referring types are singular person-referring, multiple person-referring, other person-referring, non-person-referring, and filler. Counts for each of the five referring types are shown as a group for each of the grammatical types, though in some cases counts are either zero or negligible, and the bar cannot be seen.



(1PS, 1PP, 2P, 3PS, 3PP), the quantified pronouns (QUANT), and a group including all other expressions (OTHER). Table 4.1 shows the relative frequency for the grammatical types and the most frequent expressions.

As is usually found in conversation, first-person and second-person pronouns are the most frequent, collectively comprising 82.0% of all person-referring expressions. Of particular interest, due to their high frequency and multiple possible referential meanings, are the 1PP and 2P categories (e.g., *we* and *you*), comprising respectively 24.6% and 23.7% of all person-referring expressions.

Table 4.1 also shows the information entropy of the referring type, given the gram-

Table 4.1: A statistical summary of all the markables in the dataset by grammatical type, showing their frequency relative to all markables. The entropy of the referring type given the grammatical type is also shown, as is a list of the most frequent examples of each category.

Grammatical Type	% of all Markables	Entropy (bits)	Most common words
1PS	33.7	.57	*I, my, me*
1PP	24.6	.67	*we, our, us*
2P	23.7	1.78	*you, your, yours*
3PS	.9	.66	*he, his, she*
3PP	7.2	1.25	*they, them, their*
QUANT	1.0	1.14	*everyone, everybody*
OTHER	8.9	1.57	*people, guys, user*

matical category. This measures the uncertainty one has about the type, given knowledge of only the grammatical type of the expression. The analysis reveals that second-person pronouns are a particularly challenging reference resolution problem, with a broad and relatively even distribution across referring types.

Table 4.2 lists the most commonly occurring non-person-referring personal pronouns. As expected, the 3PP pronouns are the most frequently non-person-referring, as they may be used to refer to non-human entities. Note that the 3PS pronoun *he* is commonly non-person-referring. This is because the AMI corpus scenario involves describing and drawing one's favorite animal to others. In this activity, *he* is often used to refer to the picture of the animal being drawn. Non-person-referring also occurs when pronouns are used as determiners in expressions such as *you guys*, *you both*, or *you two*, which according to the scheme requires the annotator to consider *you* as FUNC-NON-PREF and to label the subsequent word according to the referential properties of the entire phrase. Another case of non-person-referring includes personification of non-person entities, e.g., "The remote answers, '*I*'m here.'"

4.1.4 Inter-coder reliability and error analysis

To show that the produced annotations are consistent and informative, it must be established that the subjective distinctions defined in the scheme may be applied by individuals other than the scheme developers. To do this, inter-coder agreement is assessed between two independent annotators on four meetings from the AMI corpus, using Cohen's Kappa (Cohen, 1960). Each of the decisions in the annotation procedure are assessed separately:

Table 4.2: Non-person-referring (FUNC-NON-PREF) in personal pronouns, as a percentage of total occurrences of the pronoun. (Only words that are labeled at least ten times are listed.)

Word	% non-person
them	86.7
they	45.5
their	19.3
mine	6.3
he	4.5
us	3.9
you	2.4
my	1.7
I	1.4
our	1.0
me	0.8
your	0.4
we	0.1

markable identification, labeling referentiality, labeling specificity of person referents, and labeling addressing inclusion attributes. Because each decision depends on the previous, I employ a hierarchical assessment procedure that considers only instances where the annotators have agreed on previous decisions. This kind of multi-level assessment corresponds to that described and used in Carletta et al. (1997).

Markables The first annotation decision of interest is the identification of markables. Markables are either automatically identified occurrences of a pre-defined list of pronouns, or they are identified manually by the annotators. Agreement on this task, assessed only for manually identified words, was very good (κ=.94). Error analysis shows that the main issue with this decision was not determining lexical heads, but rather determining whether phrases such as "all age *groups*," "the older *generation*," and "the business *market*" should be considered as referring to *people* or not.

Person referentiality The next annotation decision is between person-referring and non-person-referring markables. For assessment of this choice, agreement is measured on a three-way categorization of the agreed markables as either FUNC-NON-PREF, FUNC-FILLER, or one of the FUNC-PREF-* categories. Agreement on this task was good (κ=.77). The only errors occurred on first- and second-person pronouns and between the FUNC-NON-PREF and

FUNC-PREF-* categories. Error analysis suggests confusion tends to occur when pronouns are used with semantically light verbs like *go*, *get*, and *have*, for example in phrases such as “there *we* go” and “*you*’ve got the main things on the front.” As in the latter example, some of the difficult choices appear to involve descriptions of states, which the speaker can choose to express either from various participants’ points of view, as above, or alternatively without explicit subjectivity, e.g., “the main things are on the front.”

Specificity and cardinality The next assessment is made on the decision between referring specifically to a single person (PERSON-SINGLE), to multiple people (PERSON-MULTIPLE), or as underspecified (also referred to as PERSON-OTHER). Agreement on this choice was very good (κ=.91), though considering only the difficult 1PP and 2P grammatical categories (e.g., *we* and *you*), agreement was less strong (κ=.75). Note that due to the hierarchical nature of the scheme, evaluation considered only cases where both annotators labeled a word as person-referring. Errors on this decision often involved ambiguities in addressing, where one annotator believed a particular individual was being addressed by *you* and the other thought the whole group was being addressed. Another common disagreement was on cases such as “*we* want it to be original,” where *we* was interpreted by one annotator as referring to the present group of participants, but by the other as (presumably) referring to the organization to which the participants belong.

Addressing inclusion attributes For the three inclusion attributes for underspecified referents (ATTR-*-INCL), agreement is calculated three times, once for each of the binary attributes. Agreement was good, though slightly problematic for addressee inclusion (speaker κ=.72; addressee κ=.50; other κ=.66). Disagreements were mainly for occurrences of *you* like the example of autobiography above. For example, “it’s *your* best friend” was used to explain why a dog is the speaker’s favorite animal, and the annotators disagreed on whether the addressee was included.

4.1.5 Summary

I have presented an annotation scheme and a set of annotations that address *participant reference*—a conversational language phenomenon that has seen little previous annotation work. The scheme focuses on addressing new distinctions concerning vagueness, ambiguity, and contextual dependency. This focus is motivated by potential applications for such distinctions, which are hypothesized to help distinguish, label, and summarize conversational activities.

Based on analysis of inter-annotator agreement, the major distinctions proposed by the scheme appear to be reliably codeable. In addition, my statistical analysis showed that the dataset contains a wide variety of participant references and should be a useful resource for several reference resolution problems for conversation. The proposed novel distinction between specific reference to real individuals and other kinds of person reference appears to be reliably codeable. The novel addressing-based distinctions for underspecified reference are less reliable but still adequate as an empirical resource for subsequent automatic processing tasks. They could also serve as reliable sources of training and test data for reference resolution experiments (though this topic will not be covered in this thesis).

4.2 Unbiased Discourse Segmentation Evaluation[3]

One of the fundamental problems in natural language processing is *discourse segmentation*—the partitioning of a stream of text, speech, or video into useful discourse-level units, such as topics in a meeting, stories in a news broadcast, or sections in a structured document. Of course, discourse segmentation is of particular importance to this thesis as it is necessary to identifying the temporal limits of an activity episode. In this section, I cover *segmentation evaluation*, an essential pre-requisite to these segmentation experiments.

It is useful to begin by distinguishing two types of segmentation task. The first type is a joint segmentation and labeling task. In this type of task, a document is segmented and each segment is assigned a label from a set of pre-defined classes. This approach has been applied most commonly in the domain of biomedical informatics (Guo et al., 2010; Hirohata et al., 2008; Teufel and Moens, 2002), where the objective is a functional, genre-specific labeling of document sections. For example, Hirohata et al. (2008) segment abstracts from scientific articles and label the produced segments according to the categories *objective*, *methods*, *results*, and *conclusions*. This type of labeling is motivated by its usefulness in information retrieval and summarization applications, where, for example, users may be more interested in conclusions than they are in methods. All such joint segmentation and labeling studies known to the author have employed evaluations in which it is the labeling that is principally relevant. Specifically, the label of each sentence is eval-

[3] A condensed version of this section was published as “Unbiased Discourse Segmentation Evaluation” in the 2010 IEEE Workshop on Spoken Language Technology, pp. 43–48, Berkeley, CA, U.S.A., December 12-10.

uated against a gold standard label, and summary results are reported using classification F-score, precision, recall, and accuracy (annotator agreement, if reported, is given using Cohen's κ).

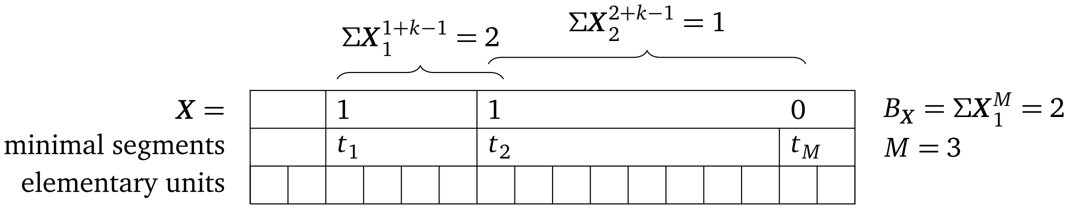
The second type of segmentation task does not involve segment labeling. Only the position of segment boundaries is considered relevant for evaluation. This type of task is most common in the literature on topic-based segmentation of text and speech, and reflects the type of segmentation experiments we shall study in this thesis. For this reason, the remainder of this chapter shall investigate *unlabeled* segmentation evaluation.

For the past decade, unlabeled segmentation studies have used $\mathrm{P_k}$ (Beeferman et al., 1999) and WindowDiff (WD) (Pevzner and Hearst, 2002) as performance measures (e.g., Eisenstein and Barzilay (2008); Utiyama and Isahara (2001); Galley et al. (2003); Malioutov and Barzilay (2006)). Instead of evaluating each segment boundary independently, $\mathrm{P_k}$ and WD accommodate near-miss boundaries by evaluating several adjacent decisions together. This provides a tolerance that is useful for segments like activity episodes, where boundaries are rare, precision placement is unreliable, and partial credit is warranted.

Recent studies, however, have noted problems with $\mathrm{P_k}$ and WD, including a biased treatment of boundaries near the beginning or end of a discourse (Sherman and Liu, 2008) and a bias related to the number of segments (Georgescul et al., 2006). Some recent publications have therefore supplemented results with segment counts (Hsueh, 2008b) or suggested novel measures (Sherman and Liu, 2008; Georgescul et al., 2006; Lamprier et al., 2007).

This study includes a novel account of the aforementioned problems, going beyond others' in three principal ways. First, I provide a rigorous analytical explanation of the measures' biases. Second, I replicate a recently published topic segmentation experiment (Eisenstein and Barzilay, 2008), showing that segmentations produced by some state-of-the-art algorithms exhibit properties whereby they benefit significantly from the biases. Third, I provide a comprehensive unbiased evaluation procedure where previous proposals have addressed only part of the problem. By replicating the topic segmentation experiment with the proposed novel measures, I come to substantially different conclusions about the efficacy of some state-of-the-art algorithms. The work thus raises questions about the validity of many previous evaluations, and it identifies appropriate algorithms and evaluation methods for use in subsequent experiments.

Figure 4.4: A graphical depiction of a segmentation and k-length window summation. Elementary units can be characters, words, seconds, etc., depending on the application. *Minimal segments* are defined by the sequence of M *potential boundaries* $\langle t_1, t_2, ..., t_M \rangle$ and represent the smallest possible segmentation. Segmentations are compared according to a sequence of summations over windows of length k, iterated and defined over minimal segments. In this example, k=2.



4.2.1 Definitions and notation

The most common type of discourse segmentation involves **linear** (non-hierarchical, contiguous) and **coarse-grained** segmentations (defined here as cases where a segment boundary occurs at fewer than half of the potential segmentation points). It is often appropriate to evaluate such segmentations in a way that gives partial-credit to approximately-correct answers, and P_k and WD have been designed to explicitly address this need. They do this by iterating a short window through the discourse from beginning to end. At each iteration, the number of boundaries contained within the window is evaluated as either correct or incorrect, with a penalty of 1 for incorrect answers. The total number of errors is then normalized by the total number of assessed windows to produce an empirical probability of error. The difference between the two measures is that P_k does not distinguish between non-zero window sums, whereas WD does. In the rest of this section, these definitions and the required notation are formalized. Some of the formalisms are depicted graphically in Figure 4.4.

I shall refer to the object of a segmentation as a **discourse**, used generically to signify any linear object such as a text or speech recording. Each discourse is associated with an M-length sequence $\langle t_1, t_2, ..., t_M \rangle$ of **potential boundaries**, which specify M+1 contiguous **minimal segments** $\langle [0, t_1), [t_1, t_2), ..., [t_M, D] \rangle$ for a discourse of length D.[4]

A **segmentation** $\boldsymbol{X}$ is defined as a sequence of Boolean variables $\langle X_1, X_2, ..., X_M \rangle$ corresponding to potential boundaries, such that X_m=1 if there is a **boundary** at t_m, and X_m=0

[4] The original definitions for P_k and WD are ambiguous as to the *values* and *units* of the potential boundaries. In practice, studies have employed a variety of specifications. This has no affect on the analytical findings presented here, though it does have a significant influence on empirical results (see Section 4.2.3).

otherwise. Let $\boldsymbol{X}_i^j$ denote a subsequence of $\boldsymbol{X}$ from i to j inclusive, referred to as a **window**, and let $\Sigma\boldsymbol{X}_i^j$ denote its sum. Let $B_{\boldsymbol{X}} = \Sigma\boldsymbol{X}_1^M$ denote the total number of boundaries in $\boldsymbol{X}$. The total number of segments is thus $B_{\boldsymbol{X}}+1$. Finally, let $\mathrm{rand}(m, b)$ denote a random permutation of an m-length segmentation with b boundaries.

I now define the two evaluation measures $\mathrm{P_k}$ and WD. Let $\boldsymbol{R}$ be a reference segmentation, and let $\boldsymbol{H}$ be an hypothesis segmentation. Let $\delta(\boldsymbol{X}, i, k)$ be the Boolean **boundary presence indicator function**, indicating whether a boundary is present in a k-length window starting at i, i.e.,

$$\delta(\boldsymbol{X}, i, k) = \begin{cases} 0 & \text{if } \Sigma\boldsymbol{X}_i^{i+k-1} = 0 \\ 1 & \text{otherwise.} \end{cases} \tag{4.1}$$

$\mathrm{P_k}(\boldsymbol{R}, \boldsymbol{H})$ can now be expressed in terms of this indicator function, such that

$$\mathrm{P_k}(\boldsymbol{R}, \boldsymbol{H}) = \frac{\sum_{i=1}^{M-k+1} \boldsymbol{1}\left[\delta(\boldsymbol{R}, i, k) - \delta(\boldsymbol{H}, i, k)\right]}{M - k + 1} \tag{4.2}$$

where k denotes the **window length**[5] and $\boldsymbol{1}[x]$ is an indicator function for non-zero numbers, evaluating to 1 if $x \neq 0$, and 0 otherwise. WD is slightly different, replacing $\mathrm{P_k}$'s boundary presence indicator function with a sum. For each window, the number of boundaries in the window is compared, and any disagreement gets a penalty of 1.[6] $\mathrm{WD}(\boldsymbol{R}, \boldsymbol{H})$ can therefore be expressed as

$$\mathrm{WD}(\boldsymbol{R}, \boldsymbol{H}) = \frac{\sum_{i=1}^{M-k+1} \boldsymbol{1}\left[\Sigma\boldsymbol{R}_i^{i+k-1} - \Sigma\boldsymbol{H}_i^{i+k-1}\right]}{M - k + 1} \tag{4.3}$$

4.2.2 Biases in $\mathrm{P_k}$ and WD

I now investigate previous suggestions that $\mathrm{P_k}$ and WD have a bias related to the number of hypothesized segments (Georgescul et al., 2006), a property I refer to as *count bias*. I also study the problem of *edge bias*, where boundaries placed at the beginning or end (i.e.,

[5] I specify k as half the average reference segment length, for which I use the formula $\max(1, \lfloor (M+1)/(2\,(B_{\boldsymbol{R}}+1)) \rfloor)$, where $\lfloor x \rfloor$ denotes the floor function. This definition is consistent with the description in Beeferman et al. (1999), though it resolves ambiguities in that original definition. These ambiguities have prompted discrepancies in the implementation of k and the windowing procedure in previous work (see discussion in Section 4.2.3).

[6] The analysis of WD in Lamprier et al. (2007) is based upon an incorrect definition where the assessed penalty is defined as the absolute value of the difference itself, rather than 1 as specified in the original definition (Pevzner and Hearst, 2002).

edges) of a discourse are counted less often (Sherman and Liu, 2008). The analysis covers two common types of segmentation experiments: (1) where the number of hypothesized segments $B_{\boldsymbol{H}}$ is unconstrained, and (2) where the experiment assumes a fixed prior $B_{\boldsymbol{H}}$. The goal here is to derive the best-scoring naive hypothesis for each experiment type. This serves the purpose of identifying the properties of segmentations that cause them to benefit from the biases.

Experiment type 1: Unconstrained $B_{\boldsymbol{H}}$

Consider a random segmentation $\boldsymbol{Y} = \text{rand}(M, b)$ of length M with b boundaries. I begin by considering the probability of the number of boundaries in a randomly selected window of $\boldsymbol{Y}$ containing k elements. This may be modeled as an urn problem without replacement, and has the distribution $\Sigma \boldsymbol{Y}_i^{i+k-1} \sim \text{Hypergeometric}(M, b, k)$. The probability of n boundaries in a randomly chosen k-length window is thus

$$P\left(\Sigma \boldsymbol{Y}_i^{i+k-1} = n\right) = \frac{\binom{k}{n}\binom{M-k}{b-n}}{\binom{M}{b}} = P_{n,k}^{\boldsymbol{Y}} \tag{4.4}$$

This probability is independent of i, so I simplify the notation as $P_{n,k}^{\boldsymbol{Y}}$. Next, given two independent random segmentations $\boldsymbol{R}$ and $\boldsymbol{H}$ representing a reference and hypothesis segmentation, I formulate the expected value of P_k's summand term (i.e., for a randomly selected window, the probability of disagreement between the segmentations), defining it in terms of the probability of zero boundaries in a window

$$\begin{aligned} &\boldsymbol{E}\left[\boldsymbol{1}\left[\delta(\boldsymbol{R}, i, k) - \delta(\boldsymbol{H}, i, k)\right]\right] \\ &\quad = P\big(\delta(\boldsymbol{R}, i, k) = 1\big) P\big(\delta(\boldsymbol{H}, i, k) = 0\big) + P\big(\delta(\boldsymbol{H}, i, k) = 1\big) P\big(\delta(\boldsymbol{R}, i, k) = 0\big) \\ &\quad = \left(1 - P_{0,k}^{\boldsymbol{R}}\right) P_{0,k}^{\boldsymbol{H}} + \left(1 - P_{0,k}^{\boldsymbol{H}}\right) P_{0,k}^{\boldsymbol{R}} \end{aligned} \tag{4.5}$$

Now, if one assumes a coarse-grained segmentation (i.e., $B_{\boldsymbol{R}} < \frac{M}{2}$), then by Equation (4.4) and the definition of the window length k, it can be shown that $P_{0,k}^{\boldsymbol{R}} > \frac{1}{2}$. Given this, the value of Equation (4.5) is monotonically decreasing as $P_{0,k}^{\boldsymbol{H}}$ increases, or equivalently, as $B_{\boldsymbol{H}}$ decreases. In other words, as the probability of a zero-sum window in the hypothesis increases, the expectation of P_k's summand decreases, thus improving P_k. One can therefore conclude the following:

Conclusion 1 In experiments with an unconstrained number of hypothesized segments, the optimal random hypothesis w.r.t. $\mathrm{P_k}$ has zero boundaries (i.e., the *null* baseline).

One can perform a similar analysis for WD. The expected value of the summand in Equation (4.3) may be computed by summing, for each possible value n of the number of boundaries in a window, the probability of that number occurring in both segmentations (i.e., obtaining a correct answer), and then subtracting this result from 1, such that

$$\boldsymbol{E}\left[\boldsymbol{1}\left[\Sigma\boldsymbol{R}_i^{i+k-1} - \Sigma\boldsymbol{H}_i^{i+k-1}\right]\right] = 1 - \sum\nolimits_{n=0}^{k}\left(P_{n,k}^{\boldsymbol{R}} \times P_{n,k}^{\boldsymbol{H}}\right) \tag{4.6}$$

Assuming $B_{\boldsymbol{R}}<\frac{M}{2}$, and given the definition of k, the value of the left side of the product is monotonically increasing as n decreases, such that n=0 has the greatest value. As each side is a probability distribution over n independently summing to 1, reducing $B_{\boldsymbol{H}}$ focuses the probability mass of the right side toward 0 and monotonically increases the product. In other words, reducing $B_{\boldsymbol{H}}$ reduces the expectation of WD's summand. One can therefore conclude the following:

Conclusion 2 The null baseline is the optimal random hypothesis w.r.t. WD.

Experiment type 2: Fixed B_H

It would appear from the previous analysis that $\mathrm{P_k}$ and WD favor a reduced number of hypothesized segments $B_{\boldsymbol{H}}$, an idea supported by the analysis in (Georgescul et al., 2006). But such a formulation is not entirely correct. Rather, it is the probability $P_{0,k}^{\boldsymbol{H}}$ which is the principal factor in *count bias*. This can be demonstrated through the design of the optimal naive hypothesis *for experiments with a fixed positive number of boundaries* $B_{\boldsymbol{H}}$>0.

The analysis above considered random segmentations, for which $P_{0,k}^{\boldsymbol{H}}$ is derived from Equation (4.4). But considering *arbitrary* segmentations and the *empirical* probability of a zero-boundary window

$$\hat{P}_{0,k}^{\boldsymbol{H}} = 1 - \frac{\sum_{i=1}^{M-k+1}\delta(\boldsymbol{H},i,k)}{M-k+1} \tag{4.7}$$

segmentations can exploit the fact that $\hat{P}_{0,k}^{\boldsymbol{H}}$ can be increased while maintaining fixed $B_{\boldsymbol{H}}$ by placing boundaries close to one another. I call this property **clumping**. Segmentations can also exploit the fact that the window iteration procedure causes boundaries at the edges of a discourse to appear in fewer windows. A segmentation can thus increase $\hat{P}_{0,k}^{\boldsymbol{H}}$ further while maintaining fixed $B_{\boldsymbol{H}}$ by moving all boundaries to the *edges* of the discourse.

I therefore propose a fixed-$B_{\boldsymbol{H}}$ baseline I call the **edge-clump baseline**. I define edge-clump as a segmentation $\boldsymbol{H} = \langle H_1, ..., H_M \rangle$ such that H_i=1 for $i \leq B_{\boldsymbol{H}}$ and H_i=0 for $i > B_{\boldsymbol{H}}$. This segmentation simply places the required $B_{\boldsymbol{H}}$ boundaries in the first $B_{\boldsymbol{H}}$ possible locations. Because no other segmentation can further increase the probability $\hat{P}_{0,k}^{\boldsymbol{H}}$ of zero boundaries in a window, one can easily arrive at the following conclusion.

Conclusion 3 For fixed-$B_{\boldsymbol{H}}$ experiments, the edge-clump baseline is the optimal naive hypothesis w.r.t. both $\mathrm{P_k}$ and WD.

Empirical confirmation and implications

The analysis above may be conveniently summarized as follows. For coarse-grained segmentations, $\mathrm{P_k}$ and WD exhibit count bias, which may be explained in terms of a majority class baseline. Since the majority of windows over the reference contain zero boundaries, the optimal naive hypothesis maximizes the number of zero-sum windows. This means that when the number of boundaries $B_{\boldsymbol{H}}$ is unconstrained, the null baseline is optimal. If $B_{\boldsymbol{H}}$ is fixed, then clumping and edge bias may be used to maximize the number of zero-sum windows.

To provide empirical confirmation of this analysis, Table 4.3 shows the results[7] of applying the null, edge-clump, and random (rand($B_{\boldsymbol{R}}$)) baselines in a replication of an experiment described in Eisenstein and Barzilay (2008), henceforth EB08.[8] The results support the analysis—the null baseline performs better than random, and edge-clump performs nearly equivalently to the null baseline. This raises questions about evaluations that use $\mathrm{P_k}$ and WD but do not consider clumping and edge bias in their analysis. It also confirms that simply reporting segment counts (Hsueh, 2008b) is inadequate for addressing the count bias problem.

[7] All results in Section 4.2 are presented as means over a collection of discourse-dependent scores. Standard deviations are given if space allows. Unless specified otherwise, I compute expected values for the random baselines and empirical values for all others. Where statistically significant differences are described, a pair-wise Student's *t*-test for $p<0.05$ is used.

[8] I provide full details of the EB08 experiment in Section 4.2.4.

Table 4.3: Results of the EB08 experiment for three different naive segmentations.

Baseline	$\mathrm{P_k}$	WD
Null	.317 (σ=.091)	.317
Edge-clump	.316 (σ=.090)	.318
Random	.427 (σ=.093)	.452

4.2.3 Definitional ambiguities

A survey of segmentation literature and publicly available evaluation software suggests there are several differing definitions of $\mathrm{P_k}$ and WD being used. The differences arise in three areas of the specification of windows: length calculation, unitization, and iteration. In this section, I show how these differences have implications for the comparability of previous experiments.

What is "average segment length?"

One ambiguity arises from the window length k being defined as "half the average reference segment length" (Beeferman et al., 1999). A literature survey shows there is inconsistency in whether multiple test discourses are concatenated into a single test document (e.g., Allan et al. (1998)) or treated separately (e.g., Eisenstein and Barzilay (2008)). This means there are three possible methods for computing k: (A) it may be computed separately for each discourse, (B) the mean of these discourse-dependent means may be used, or (C) a single corpus-wide mean of segment lengths may be computed.

Table 4.4 shows the results of the edge-clump baseline on the replicated EB08 experiment, applying each of the three methods of computing k. The choice of method has considerable effects on scores (and their variance). For $\mathrm{P_k}$, the results using method A are 25% lower than method B. This raises questions of comparability for experiments where the method is unspecified.[9]

[9] The definition of k used in this paper is a variant of method A. This choice is based upon its common use in recent work. Since k must be a positive integer, my definition uses the max and floor functions. It was confirmed by examination of public software that my definition corresponds to Malioutov and Barzilay (2006) and Eisenstein and Barzilay (2008), though others use the round function instead (Choi, 2000; Lamprier et al., 2007). The TDT evaluations (NIST, 1998) use method C.

Table 4.4: Results for edge-clump using three different methods of computing k.

Definition of k	P_k	WD
A	.316 (σ=.090)	.318
B	.417 (σ=.225)	.420
C	.407 (σ=.223)	.410

Units, window iteration, and granularity

In addition to specifying a method for calculating average segment length, disambiguating the measures also requires specification of three other properties: (1) the units of k, (2) the set of potential boundaries, and (3) a method of window iteration. In my definition, the unitization and iteration of windows is specified formally in relation to the potential boundaries. However, Beeferman et al. (1999) are ambiguous about the relationship of these three factors.

Concerning unitization, Beeferman et al. (1999) suggest measuring k in words. However, using words is clearly not applicable to all segmentation tasks. Concerning iteration, it is unclear whether a window (say, in word units) is iterated over every word or every potential boundary (say, sentence units). A literature survey suggests a variety of methods are used.

Assuming that unitization and iteration are specified according to the definition given in this thesis, the method for establishing potential boundaries may still vary. For example, a literature survey shows that experiments on topic segmentation of meetings have used differing methods despite employing the same reference segmentation. Galley et al. (2003), for example, use what they call "speaker changes" as potential boundaries. These appear to be equivalent to the start times of "time bins" identified within the ICSI corpus.[10] Eisenstein and Barzilay (2008) use a segmentation derived from this one, in which the time bins are further divided where transcribed sentence endings occur. Hsueh (2008b) uses "spurts," units produced by an automatic silence detector.

Using the same reference segmentation as above, the edge-clump baseline was applied in an experiment comparing two choices for minimal segment: the start times of the 'time bins' in the ICSI corpus `.mrt` files (MRT), and the start times of annotated dialogue acts (Shriberg et al., 2004) (MRDA). The results of this experiment are given in Table 4.5.

[10] "Time bin" units are an atheoretical notion that is specific to the ICSI corpus. They are based on the output of a speech-non-speech detector, arbitrarily modified by transcribers for convenience of transcription (see corpus documentation).

Table 4.5: The effects on P_k and WD of specifying minimal segments.

Minimal Segment	P_k	WD
MRT	.330 (σ=.084)	.333
MRDA	.293 (σ=.092)	.296

The P_k score for method MRDA is 11% lower than for method MRT. This shows that the granularity and distribution of minimal segments can have a dramatic effect on scores.

Implications and solutions

The literature survey and empirical tests presented in this section raise serious questions about the replicability and comparability of previous experiments. To resolve these issues, I propose that future segmentation evaluations that use windowed measures require the following:

Requirement 1 The method for computing the length of the window must be unambiguous.

Requirement 2 The units and iteration procedures for windows must be unambiguous.

Requirement 3 An unambiguous definition of potential boundaries must be specified.

Importantly, I do not propose one single solution for these requirements. While any ambiguity must be eliminated, the flexibility presented in these choices is a benefit. First, it allows the measures to be applied to new corpora and tasks. Second, varying window length allows changes in the level of tolerance of near misses.

4.2.4 An empirical re-evaluation

I now investigate how much the measures' biases affect evaluation of segmentations produced by state-of-the-art algorithms. To do this, I replicate and expand the EB08 experiment comparing unsupervised lexical-semantic topic segmentation algorithms: BayesSeg (Eisenstein and Barzilay, 2008), LCseg (Galley et al., 2003), MinCutSeg (Malioutov and

Table 4.6: Results from the replicated EB08 experiment (original $\mathrm{P_k}$ score in *italics*).

Algorithm	**$\mathrm{P_k}$**	EB08	**WD**
BayesSeg	.272 (σ=.112)	*.264*	.324
C99	.384 (σ=.135)		.408
LCseg	.308 (σ=.116)	*.309*	.322
MinCutSeg	.347 (σ=.138)	*.370*	.363
U00	.312 (σ=.127)	*.297*	.344
Null	.317 (σ=.091)		.317
Edge-clump	.316 (σ=.090)		.318
Random	.427 (σ=.093)		.452

Barzilay, 2006), and U00 (Utiyama and Isahara, 2001), to which I add C99 (Choi, 2000).[11]

The data comprise topic segmentations (Galley et al., 2003) of 25 meetings from the ICSI meeting corpus (Janin et al., 2003). These data are pre-processed by a script (Eisenstein and Barzilay, 2008) which removes two types of segments that are unique to the corpus: those containing recited digits (used by the corpus designers for a speech recognition task), and those segments at the beginning of the meetings where seating or preparation of the recording takes place. This results in a total of 112 segment boundaries for the corpus. The script produces minimal segments based upon transcription units in the corpus `.mrt` files, though it also adds any unit-internal sentence boundaries as potential boundaries. The experiment is a fixed-B_H experiment, where the algorithms are given a known number of segments B_H=B_R, derived from the human-annotated reference.

The results are shown in Table 4.6. (Note there are differences with the originally reported results, attributable to the use of independent evaluation software.) All the algorithms except C99 are significantly better than the random baseline for both $\mathrm{P_k}$ and WD. Only BayesSeg's $\mathrm{P_k}$ score is significantly better than edge-clump.

These results are clearly problematic. The edge-clump score is much better than random, confirming the potential influence of clumping and edge bias. Since it has not been confirmed whether the algorithms' output exhibits these properties as well, and since most of the scores are not significantly better than edge-clump, the question arises whether these algorithms are indeed better than a naive baseline. The next aim must therefore be to estimate how much of the algorithms' results are attributable to clumping and edge bias and

[11] All algorithms are run using default parameter settings (or if provided, the recommended settings for use on this dataset). Evaluation is performed using software developed specifically for this study.

how much is due to accurate boundary placement. To do this, I develop novel statistics and an experiment involving random reference segmentations.

Statistics for clumping and edge bias

I propose two statistics that are useful for identifying clumping: $\sigma(\boldsymbol{L})$ and $\sigma(\frac{1}{\boldsymbol{L}})$. Their calculation proceeds as follows. For each hypothesis segmentation $\boldsymbol{H}$, obtain an ordered $B_{\boldsymbol{H}}$-length sequence of *normalized boundary indices* $\boldsymbol{I} = \langle I_1, ..., I_{B_{\boldsymbol{H}}} \rangle = \langle i : H_i{=}1 \rangle \times \frac{1}{M+1}$ from which the $B_{\boldsymbol{H}}{+}1$-length sequence of *normalized segment lengths* $\boldsymbol{L} = \langle I_1, I_2 - I_1, ..., M{+}1 - I_{B_{\boldsymbol{H}}} \rangle$ can be computed, where M is the number of possible boundaries. The proposed statistics are the standard deviation of the normalized segment lengths $\sigma(\boldsymbol{L})$ and the standard deviation of their inverses $\sigma(\frac{1}{\boldsymbol{L}})$. Both of these measure the evenness of a segmentation, where a perfectly even segmentation (i.e., all segments are the same length) measures 0, and greater numbers indicate increased clumping. $\sigma(\boldsymbol{L})$ highlights the occurrence of long segments, while $\sigma(\frac{1}{\boldsymbol{L}})$ highlights short segments.

To measure edge bias, I propose the statistic $\mu(\Sigma_k)$, which calculates the mean number of segment boundaries in a window $\sum_{i=1}^{M-k+1} \Sigma \boldsymbol{H}_i^{i+k-1} \frac{1}{M-k+1}$, where lower values indicate more edge bias.

I also propose a statistic for measuring *index bias* called $\sigma(\boldsymbol{I})$, which measures the tendency for boundaries to occur in some locations more often than others. For this, one calculates the standard deviation $\sigma(\boldsymbol{I})$ of the values of the binned frequency distribution (i.e., histogram) of the normalized boundary indices $\boldsymbol{I}$ for the entire corpus (I use a 40-bin histogram in this paper).

For each statistic presented here (other than $\boldsymbol{I}$), a value for each discourse is computed and the mean of those values is provided.

Measuring the influence on $\mathrm{P_k}$ and WD

In addition to statistics, one can also directly measure the effect that clumping and edge bias have on scores. One can evaluate each hypothesis against a *random reference* segmentation. One can then report the difference, denoted $\Delta_{\mathrm{P_k}}^{\mathrm{rand}}$, between this score and the score obtained by a random hypothesis. Lack of a difference is an indication that clumping and edge bias are not a factor. A negative difference indicates they are benefiting the result while a positive difference indicates the opposite (i.e., that even spacing or a lack of edge bias is hurting the score).

Table 4.7: Clumping (segment length) statistics $\sigma(\boldsymbol{L})$ and $\sigma(\frac{1}{\boldsymbol{L}})$, edge bias statistic $\mu(\Sigma_k)$, index bias statistic $\sigma(\boldsymbol{I})$, and random reference experiment results $\Delta^{\text{rand}}_{\mathbf{P_k}}$. For the random baseline, statistics are computed over a set of generated examples.

Algorithm	$\sigma(\boldsymbol{L})$	$\sigma(\frac{1}{\boldsymbol{L}})$	$\mu(\Sigma_k)$	$\sigma(\boldsymbol{I})$	$\Delta^{\text{rand}}_{\mathbf{P_k}}$
BayesSeg	.162	364	.381	4.7	−.014
C99	.147	6	.420	2.5	+.018
LCseg	.361	203	.106	18.3	−.076
MinCutSeg	.100	94	.404	3.3	+.015
U00	.095	12	.422	4.5	+.018
Edge-clump	.438	648	.011	35.4	−.102
Random	.172	33	.396	4.3	

Table 4.7 shows the results of performing such an analysis. For conciseness, only $\mathrm{P_k}$ scores for the random reference experiment are shown, though it has been confirmed that the effect on WD is similar. As expected, the statistical measures are consistent with the results of the random reference experiment. The results indicate that LCseg and BayesSeg produce clumped boundaries, and that their scores benefit from this. The others' appear to produce more evenly-spaced boundaries, which hurts their scores.

One notable result is that BayesSeg seems to use extreme clumping as indicated by $\sigma(\frac{1}{\boldsymbol{L}})$, but with less benefit to its scores than LCseg. While the results suggest this is due to lack of edge bias in BayesSeg, to understand this further one can employ a histogram. Figure 4.5 shows the distribution of normalized segment lengths $\boldsymbol{L}$ produced by BayesSeg. The histogram shows that BayesSeg produces a high density of extremely short segments. However, there are clearly several non-clumped segments, which explains why $\sigma(\boldsymbol{L})$ is not high. BayesSeg seems to effectively divide the discourse into fewer than the specified number of segments, making up the difference by placing any remaining required boundaries immediately adjacent to the others. Its good score suggests it is acting like a high-precision, low-recall boundary detector.

Another notable result is that LCseg exhibits extreme edge bias. Figure 4.6 makes this even more clear by showing the distribution of normalized boundary indices $\boldsymbol{I}$ for LCseg. Its segmentations appear to be very similar to the edge-clump baseline which suggests that LCseg achieves much of its good score purely by virtue of clumping and edge bias.

Figure 4.5: A histogram of the normalized segment lengths $\boldsymbol{L}$ hypothesized by BayesSeg.

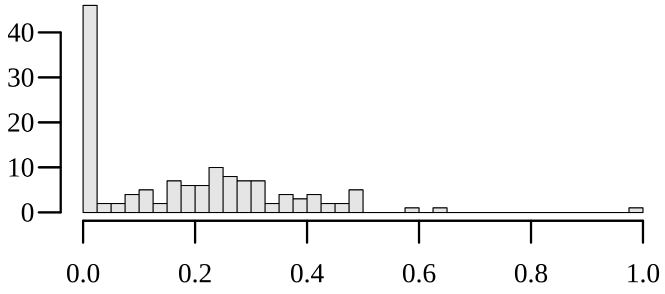
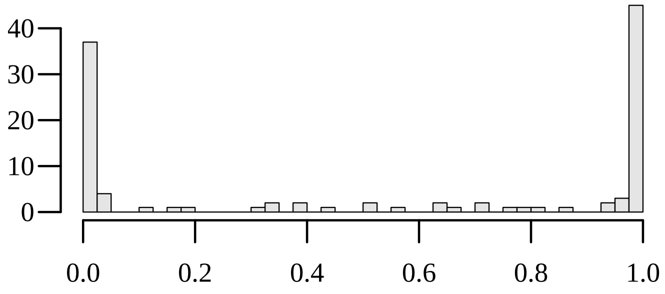


Figure 4.6: A histogram of the normalized boundary indices $\boldsymbol{I}$ hypothesized by LC-seg.



4.2.5 The proposed measures k-κ, k-precision, and k-recall

It is helpful to recall that $\mathrm{P_k}$ and WD count the number of boundaries within each window, comparing the reference to the hypothesis. The sign of the comparison term reflects one of two possible types of error. A negative value indicates a false positive boundary, a.k.a., **false alarm**, and a positive value indicates a false negative boundary, a.k.a., **miss**. $\mathrm{P_k}$ and WD simply sum the false alarms and misses, and then normalize by the total number of windows.

The analysis thus far has shown that since the opportunity of a miss is inherently less probable due to the setting of k, hypotheses improve by reducing their number of zero-sum windows. To address this, several modified versions of $\mathrm{P_k}$ and WD have been proposed. A variant of $\mathrm{P_k}$ was used in the TDT pilot evaluation (Allan et al., 1998) that weighted error types independently according to the *a priori* probability of a zero-boundary window in the reference. A similar version called C_{seg} was used in the TDT2 evaluations that instead used a WD-like boundary count distinction (NIST, 1998). Finally, a modification of WD was recently proposed (Georgescul et al., 2006). The idea that all of these proposals share is the independent normalization of error types. All of them, however, remain problematic. The Georgescul and TDT2 measures, employing WD's count-dependent error function, gives no partial credit to multiple adjacent boundaries, even when they are very close to the correct location. (Note that BayesSeg, which produces such segmentations, has a much greater difference between $\mathrm{P_k}$ and WD in Table 4.6.)[12] Both TDT alternatives, because they employ no normalization with respect to the hypothesis, are unable to treat random hypotheses equally. These measures are biased to favor, for random segmentations, those with the same proportion of zero-sum windows.

Other alternatives include a detection-error-tradeoff (DET) curve, which requires a range of miss-to-false-alarm ratios at various sensitivity settings. One can also use points on that curve, such as the equal error ratio (EER). These methods, however, are only applicable when sensitivity is adjustable, and this is not the case with many segmenters.

One possible solution is to calculate a windowed F_1 measure k-F_1. The problem is that such a measure would exhibit count bias. By performing an analysis similar to that performed for $\mathrm{P_k}$ and WD, it can be shown that for random hypotheses, the one which proposes a boundary at every location will achieve the best k-F_1 score. k-F_1 is thus biased in a manner opposite to $\mathrm{P_k}$ and WD, favoring evenly-spaced segmentations when $B_{\boldsymbol{H}}$ is fixed and a greater number of segments when $B_{\boldsymbol{H}}$ is unconstrained.

An unbiased summary measure k-κ

What is needed is a summary measure that evaluates a single hypothesis, one that is *unbiased* toward naive hypotheses, and one that gives partial credit where appropriate. For this purpose, I propose the novel measure k-κ (k-kappa), which is simply $\mathrm{P_k}$, but explicitly corrected for chance agreement *between windows*, thus ameliorating the root source of count bias. (The measure is inverted as well, so that better scores are greater than worse scores).

$$k\text{-}\kappa(\boldsymbol{R},\boldsymbol{H}) = \frac{1-\mathrm{P_k}(\boldsymbol{R},\boldsymbol{H})-C}{1-C} \tag{4.8}$$

where chance agreement C is defined

$$C = \hat{P}^{\boldsymbol{R}}_{0,k}\hat{P}^{\boldsymbol{H}}_{0,k} + (1-\hat{P}^{\boldsymbol{R}}_{0,k})(1-\hat{P}^{\boldsymbol{H}}_{0,k}) \tag{4.9}$$

[12] Note that the argument supporting WD is problematic. In Section 2.1, Pevzner and Hearst (Pevzner and Hearst, 2002) suggest that $\mathrm{P_k}$ penalizes false negatives more than false positives. This argument is invalidated by switching hypothesis and reference in their examples.

Cohen's κ (Cohen, 1960) is commonly used for measuring agreement between annotators on categorical data. The benefit of applying κ in this way is that it explicitly factors out chance agreement *between windows*. This means that segmentations are evaluated in a way that is robust to variation in the number of boundaries as well as clumping. This allows application in experiments with either fixed or unconstrained B_H (experiments for which B_H is unconstrained should therefore report B_H). For k-κ, a score of 0 reflects chance agreement, 1 is perfect agreement, and -1 is perfect disagreement (perfect disagreement is only possible when $k = 1$).

It should be noted that this does not solve the problem of edge bias. For this, one can borrow the Sherman and Liu (2008) extended windowing procedure. This involves appending 0's to each end of a segmentation so that each original boundary appears in the same number of windows. Formally, each segmentation $\boldsymbol{X}$ of length M is extended into a sequence $\boldsymbol{X}'$ of length $M{+}2k{-}2$ such that $X'_i{=}0$ for $i{<}k$ or $i{>}M{+}k{-}1$ and $X'_{i+k-1}{=}X_i$ for $1{\leq}i{\leq}M$. This extended segmentation should be used as the input to the recommended measures.

k-precision and k-recall

An adequately unbiased summary measure is useful for making broad claims about segmentation performance, but it should not be used alone. An evaluation should also give an independent account of performance in relation to *both types of error* (i.e., false positives and false negatives). I therefore propose that evaluations also use windowed variants of precision and recall, which I call k-precision and k-recall. While using miss and false alarm probabilities would satisfy our need as well (recall is in fact the inverse of the probability of a miss), precision and recall are more widely used and well-understood. (The clumping, edge, and index bias statistics are also useful in further characterizing segmentations).

The k-precision and k-recall measures are calculated by computing, for each extended segmentation $\boldsymbol{X}'$, a sequence $\boldsymbol{W}_{\boldsymbol{X}'} = \langle W_1, W_2, ..., W_{M+k-1} \rangle$ such that $W_i = \delta(\boldsymbol{X}', i, k)$. This is the sequence of Booleans resulting from applying the boundary presence indicator function to each window. The precision and recall of a positive value (i.e., a boundary within the window) is then calculated in the usual way according to a comparison of $\boldsymbol{W}_{\boldsymbol{H}'}$ against $\boldsymbol{W}_{\boldsymbol{R}'}$.[13]

[13] Note that varying k is acceptable with the proposed measures and allows for the adjustment of the level of near-miss tolerance. While the results here use a definition of k consistent with that used by $\mathrm{P_k}$, using an application-warranted constant is likely a more informative choice. Evaluation across several choices can

Table 4.8: Results of the EB08 experiment; k-precision, k-recall, and k-κ.

Algorithm	k-prec	k-rec	k-κ
BayesSeg	.543	.437	.289
C99	.396	.414	.125
LCseg	.415	.305	.124
MinCutSeg	.451	.427	.169
U00	.486	.496	.262
Edge-clump	.449	.121	.063
Random	.383	.317	.000

Table 4.8 shows the results of applying the novel measures in the EB08 experiment. The results are consistent with the findings from prior analysis and suggest novel, more informative conclusions about algorithm performance. The main difference is that LCseg shows a severely diminished performance, and (along with C99) is not significantly better than random in terms of k-κ. Another notable result is the difference between precision and recall for BayesSeg. The analysis of BayesSeg above suggested that it performs as a high-precision classifier, and these results support this conclusion. In terms of k-κ, BayesSeg and U00 are the only algorithms that are significantly better than all others, though the difference between the two is not significant.

4.2.6 Summary

My analysis of $\mathrm{P_k}$ and WD has shown that both favor segmentations with fewer segments, clumping, and/or edge bias, while disfavoring those with more segments and/or even spacing. Backed by an empirical demonstration that these biases have a major impact on results for state-of-the-art topic segmentation algorithms, this raises serious questions about whether previously published experimental results accurately reflect algorithm effectiveness.

In response, I have proposed k-κ—an unbiased evaluation measure which corrects for chance agreement between windows. I have proposed k-precision and k-recall as near-miss tolerant correlates to precision and recall. And I have suggested several useful statistical measures of segmentation properties. Provided that the method for unitization, iteration, and computation of the window and its length k is unambiguously defined, these measures

highlight nuances in the “nearness” of different hypotheses. The proposed measures are also a generalization of those used in non-tolerant segmentation evaluations, i.e., precision and recall, when k=1.

provide a unified yet flexible framework for evaluating segmentations with and without near-miss tolerance, and with and without constraints on the number of hypothesized segments.

Chapter 5

Participant-Relational Activity Segmentation and Labeling

In this chapter, I apply participant-relational analysis experimentally to two computational problems: automatic discourse segmentation and automatic discourse segment labeling. The first set of experiments test whether participant-relational features can serve as a basis for automatically segmenting conversations into discourse segments, e.g., activity episodes. Results show that they are effective across different levels of segmentation and different corpora, and indeed sometimes more effective than the commonly-used method of using semantic links between content words, i.e., lexical cohesion. They also show that feature performance is highly dependent on segment type, suggesting that human-annotated "topic segments" are in fact a multi-dimensional, heterogeneous collection of topic and activity-oriented units. The second set of experiments concern the use of participant-relational features to automatically identify labels for discourse segments. In contrast to assigning semantic topic labels, such as topical headlines, the proposed algorithm automatically labels segments according to activity type, e.g., presentation, discussion, and evaluation. The method is unsupervised and does not learn from annotated ground truth labels. Rather, it induces the labels through correlations between discourse segment boundaries and the occurrence of bracketing meta-discourse, i.e., occasions when the participants talk explicitly about what has just occurred or what is about to occur. Results show that bracketing meta-discourse is an effective basis for identifying some labels automatically, but that its use is limited if analysis of corpus-wide correlations to segment features are not employed.

5.1 Outline of the Experiments

In Chapter 3, I identified the following tasks that must be addressed by an activity-oriented summarizer:

Task 1 Locate the main communicative activities within the dialogue.

Task 2 Recognize each activity's type, subject matter, and participation structure.

Task 3 Create a natural language description for each activity.

These three tasks are an embodiment of Spärck-Jones' three main summarization steps (see Figure 2.2): interpretation, transformation, and generation. Together, they constitute a complete activity-oriented conversational summarization system. Crafting such a system, due to the great complexity involved, is out of reach of this single thesis. Instead, this thesis aims to take important steps toward addressing specific parts of these three tasks. Namely, this thesis addresses two elementary computational problems underlying the construction of an activity-based conversation summarizer: *activity-oriented segmentation* and *activity-oriented segment labeling*.

The first problem—segmentation—is clearly the most important problem underlying Task 1. Segmentation identifies the temporal boundaries of discourse units, and thus supports locating the main communicative activities within a dialogue. The second problem—segment labeling—addresses part of Task 2, but is not a complete answer to it. Instead, segment labeling addresses the "activity type" component of the PAS structure. The other two components—"subject matter" and "participation structure"—are not covered here. Subject matter is not covered because it has been the subject of numerous previous experiments, and our goal in this thesis is to propose a technique that *complements* subject matter summarization. Participation structure is not covered due to its significant complexity. Addressing its challenges would pose too large of a problem for one thesis.

In studying our two chosen technical problems, there are a number of experimental paths that could be followed. In particular, there are four primary dimensions along which parameters of an experimental design must be considered. The first dimension concerns the *segmentation method*, i.e., the algorithm used to perform the segmentation. The second dimension is that of *segmentation basis*, e.g., whether we endeavor to produce topic-oriented or activity-oriented segments. The third dimension is that of *corpus type*, e.g., the source of the data and type of linguistic interaction being studied. Finally, the fourth dimension concerns the *features used*, e.g., whether content words or participant-relational features are used as input to the system.

With the particular goals of this thesis in mind, I shall now map out the experiments to follow, identifying where they sit within each of these four dimensions and explaining

Table 5.1: A summary of the sequence of six experiments presented in this chapter. The "Problem Type" column refers to the basic nature of the experiment, whether segmentation or segment labeling. The remaining four columns describe the main dimensions of the experimental design.

	Problem Type	Method	Basis	Corpus	Features
1	Segmentation	Purpose-built, NP co-reference	intentional	Pear	participant-relational, coref
2	Segmentation	Purpose-built, statistical	intentional, topic	Pear	participant-relational, content words
3	Segmentation	statistical	activity, topic	AMI	participant-relational, content words
4	Segmentation	statistical	activity	AMI	participant-relational (enhanced)
5	Labeling	corpus-level	activity	AMI	metadiscourse
6	Labeling	segment-level	activity	AMI	metadiscourse

why the particular progression was chosen. In total, six experiments will be presented. They are sequenced in a manner whereby each one builds upon the results of the previous and takes a step toward the ultimate goal of activity-based segmentation and labeling. The sequence is summarized in Table 5.1. Each experiment is shown as a row in the table. The "Problem Type" column refers to our two main technical problems—segmentation and segment labeling. The remaining four columns correspond to the four dimensions just described.

Experiment 1. *Can participant-relational features be used as the basis for intentionally-based discourse segmentation?* In Experiment 1, I develop a purpose-built, participant-relational segmentation algorithm called NM09.[1] The algorithm is designed specifically for fine-grained intentional segmentation of the Pear Stories,[2] a corpus of conversational monologues (Chafe, 1980b) that have been segmented using an intentional criterion (Passonneau and Litman, 1997). The basic idea behind this experiment is to test the participant-relational approach in a highly simplified scenario of intentionally-based segmentation. The algorithm is purpose-built so that its mechanism is highly controlled. The corpus was chosen because it was the subject of previous study and is the only known

[1] The NM09 name derives from its original publication in Niekrasz and Moore (2009).

[2] This corpus was briefly introduced in Section 3.6, and will be described further below

corpus of intentional segmentations. We propose that intentional segments are broadly equivalent to the activity-oriented episodes that are the main target in this thesis. Therefore, participant-relational features are hypothesized to be useful in performing such a segmentation. The experiment is designed to be firmly grounded in prior work by replicating a notable previous experiment (Passonneau and Litman, 1997), and comparing the participant-relational algorithm to Passonneau and Litman's coreference based NP algorithm. The results help to establish the general viability of the participant-relational approach to segmentation.

Experiment 2. *How does the participant-relational approach compare to state-of-the-art topic-oriented methods?* In Experiment 2, I expand on Experiment 1, using the same corpus and algorithm, but comparing the performance of NM09 to a collection of state-of-the-art segmentation algorithms that employ the principle of lexical cohesion as a fundamental motivation. Here, the idea is to begin understanding the relationship between activity-based coherence and more commonly-studied indicators of coherence, such as lexical cohesion. Comparing the results of these two essentially orthogonal methods (they use totally distinct features as input, i.e., content words vs. participant-relational features), allows us to begin the task of understanding how each of them relates independently to both the intentional and topical segmentation problems. It also allows us to establish how well the novel participant-relational method performs in relation to current state-of-the-art algorithms.

Experiments 3 and 4. *Can participant-relational features be used as the basis for coarse-grained segmentation of workplace meetings, and how is performance dependent upon segment type?* In these experiments, I begin addressing the segmentation problem in the context of the AMI corpus of multi-party meetings (Carletta, 2007), employing only the more robust statistical segmentation algorithms. I therefore no longer use the NM09 algorithm, which is specifically designed for the Pear Stories. Rather, I re-purpose existing state-of-the-art (unsupervised) statistical topic segmentation methods (Eisenstein and Barzilay, 2008; Utiyama and Isahara, 2001) so that they use participant-relational features (e.g., subjectivity, the amount of participant speech, modality, personal pronouns, etc.) as input. I then compare the performance of these re-purposed versions with their original incarnation (which uses only content word stems as input). This allows an investigation of the effect of various features as input (which is the main purpose of the study) without there being any effects from using different algorithms. Also important to this experiment, I conduct a study of the effect of using different features *dependent on the segment label*. That is, I compute how well each input feature performs for each of the several segment labels used in the AMI corpus annotations. This label-dependent investigation reveals striking patterns in the relationship between topics, activities, lexical-semantic features, and participant-relational features. These patterns, though often discussed in a qualitative setting, have not been previously explored from the statistical, corpus-oriented perspective taken here. The experiment thus represents the end-point for segmentation studies in this thesis, but the beginning of an interesting future line of corpus research addressing the multi-dimensional, heterogeneous nature of discourse structure.

Experiment 5. *Can bracketing meta-discourse be used to infer a global set of activity type labels appropriate for a corpus?* This experiment is intended to determine whether the use of communicative activity words (i.e., presentation, discuss, evaluation, etc.) correlate with activity segment boundaries. It is hypothesized that participants use meta-discourse (both before and after activity segments) that contains words that describe the type of the adjacent activity. It is also hypothesized that this occurs with enough regularity that statistical correlations (in this case, assessed using the k-κ segmentation evaluation measure) between segment boundaries and the use of such words can allow for the unsupervised extraction of a set of activity type labels for an entire corpus, i.e., at a *corpus-level* (see Table 5.1). The output of this experiment is an automatically-generated ranked list of words that are hypothesized to describe the activity types present in the corpus.

Experiment 6. *Do inferred types of activity correlate reliably enough with meta-discourse to allow such a correspondence to be used to label activity segments individually?* Whereas Experiment 5 examines whether a set of global labels can be extracted for a corpus, Experiment 6 investigates the activity-type labeling of individual activity segments, i.e., at the *segment-level* (see Table 5.1). The fundamental difference between this experiment and the previous one is that rather than identifying global correlations between meta-discourse and segment boundaries through the whole corpus, correlations are assessed on a per-segment basis. This makes the task much harder, since (using the selected approach) meta-discourse must occur at a segment's boundary for such a correlation to be discovered. While the experiment does not employ learned global correlations between participant-relational features and meta-discourse (this is reserved for future work), it does allow one to test whether meta-discourse does indeed occur, and provides an initial assessment of how helpful it might be for labeling individual segments.

5.2 Experiment 1: Fine-grained Intentional Segmentation of the Pear Stories[3]

Participant-relational analysis (described in detail in Section 3.6) proposes that language expressing participants' relationships to the dialogue is a principal indicator of the intentional structure of discourse. This implies that such language may be used as a means for performing intention-based *discourse segmentation*. In this section, I describe an initial experiment in a series of experiments that test this idea formally. I hypothesize that a simple set of participant-relational features can be used to perform automatic intentional segmentation of a corpus of conversational monologues.

The hypothesis is based upon the following reasoning. First, it is proposed that the discourse segments in the test corpus, by virtue of being intention-based, are broadly equivalent to the activity episodes that are the focus of this thesis. Then, according to the principles of participant-relational analysis, it is proposed that these segments are distinguishable by participant-relational features. If one can reliably extract such features, then an automated segmentation algorithm can be applied to them to produce an intentional segmentation.

In addition to this basic test of the participant-relational approach, the current experiment is also designed to establish a grounding in previous work. To achieve this, the experiment replicates a well-known segmentation study (Passonneau and Litman, 1997) (henceforth P&L) that examined performance on a set of conversational monologues called the Pear Stories (Chafe, 1980a) (introduced in Section 3.6.2). While the overarching goal in this thesis is to apply the participant-relational approach to workplace meetings and to a coarser-grained activity-oriented analysis, using this dataset and replicating the previous experiment serves as a useful stepping-stone toward that end. First, its results can be compared directly to previous work in the literature. Second, it tests empirically whether the participant-relational approach is indeed appropriate for *intentional* segmentation.

5.2.1 Data

The experiment uses the same dataset as P&L, a corpus of 20 spoken narrative monologues known as the Pear Stories (Chafe, 1980b) (see Section 3.6.2 for an introduction). We note

[3] A version of this section was published as "Participant Subjectivity and Involvement as a Basis for Discourse Segmentation" in the Proceedings of SIGDIAL 2009: the 10th Annual Meeting of the Special Interest Group in Discourse and Dialogue, pp. 54–61, Queen Mary University of London, September 2009.

here that the narratives have a mean of 100 prosodic phrases each and a mean 6.7 words per phrase. The experiment also uses the segmentation annotations of P&L. We note here that P&L concluded that an appropriate gold standard segmentation could be produced by using the set of boundaries assigned by at least three of the seven annotators. This is the gold standard used in this experiment. It assigns a boundary at a mean 16.9% of the possible boundary sites in each narrative, resulting in a mean discourse segment length of 5.9 prosodic phrases (or 39.5 words).[4]

5.2.2 The nm09 Algorithm

For the experiment, a novel participant-relational segmentation algorithm was designed (which I shall refer to as NM09). It was designed to be simple and easily re-implementable, though its approach is specific to the Pear Stories corpus. It should therefore not be considered as a generic, robust solution, but rather one which is designed principally for this initial study. The algorithm is composed of four main phases: pre-processing, feature extraction, similarity measurement, and boundary assignment.

Phase 1. Pre-processing

The pre-processing phase involves the use of a small number of publicly available natural language processing tools. The first pre-processing step removes disfluencies in the transcript by deleting repeated sequences of words and incomplete words. The transcript is then parsed (Klein and Manning, 2002), and a collection of typed grammatical dependencies are generated (de Marneffe et al., 2006b). Empirical evaluations of these processing systems show that parsing achieves an 86.7% F-score on the WSJ section of the Penn Treebank. The grammatical dependency generator has not been intrinsically evaluated, though its output has been used previously in a textual entailment system (de Marneffe et al., 2006a) that achieved a 60% accuracy on that challenging task. The TTT2 chunker (Grover and Tobin, 2006) is then used to perform tense and aspect tagging of verb phrase chunks. An evaluation of the chunker on the CoNLL 2000 chunking test set achieved a 91.3% F-score for verb chunks, though the tense and aspect tagging of those chunks used in this study was not evaluated. Discussion of the potential knock-on effects introduced by errors in these pre-processing systems is presented in the discussion sections of the experiments.

[4] The author thanks Becky Passonneau and Diane Litman, who kindly allowed their annotations to be used in this experiment.

Examples of the results of these processing steps as applied to **(13)** are shown below. Extract **(14)** shows the generated typed dependencies (see de Marneffe et al. (2006b) for further explanation of the dependencies). Extract **(15)** shows the tense and aspect tags generated by TTT2.

(13) [PEAR–09–21.2]
21.2 okay.
22.1 Meanwhile,
22.2 there are three little boys,
22.3 up on the road a little bit,
22.4 and they see this little accident.
23.1 And u-h they come over,
23.2 and they help him,
23.3 and you know,
23.4 help him pick up the pears and everything.
24.1 A-nd the one thing that struck me about the- three little boys that were there,
24.2 is that one had ay uh I don't know what you call them,
24.3 but it's a paddle,
24.4 and a ball-,
24.5 is attached to the paddle,
24.6 and you know you bounce it?
25.1 And that sound was really prominent.
26.1 Well anyway,
26.2 so- u-m tsk all the pears are picked up,
26.3 and he's on his way again,

(14) [PEAR–09–21.2–dependencies]
21.2
22.1
22.2 `expl(are, there), num(boys, three), amod(boys, little)...`
22.3 `dep(up, on), det(road, the), pobj(on, road), dep(little, a)...`
22.4 `dep(and, they), rcmod(they, see), det(accident, this)...`
23.1 `dep(and, uh), nsubj(come, they), rcmod(uh, come)...`
23.2 `dep(and, they), rcmod(they, help), dobj(help, him)`
23.3 `dep(and, you), rcmod(you, know)`

(14) [cont. . .]

```
23.4   nsubj(pick, him), ccomp(help, pick), prt(pick, up)...
24.1   det(thing, the), num(thing, one), dep(and, thing)...
24.2   dep(is, that), dep(is, one), aux(ay, had), rcmod(one, ay)...
24.3   cc(paddle, but), nsubj(paddle, it), cop(paddle, ’s)...
24.4   dep(and, a), pobj(a, ball)
24.5   auxpass(attached, is), prep(attached, to), det(paddle, the)...
24.6   dep(and, you), rcmod(you, know), nsubj(bounce, you)...
25.1   det(sound, that), dep(and, sound), cop(prominent, was)...
26.1   dep(well, anyway)
26.2   mark(tsk, so), nsubj(tsk, um), predet(pears, all)...
26.3   dep(and, he), rcmod(he, ’s), prep(’s, on), poss(way, his)...
```

(15) [PEAR–09–21.2–TTT2]

```
21.2   (p ((rg meanwhile)))
22.1   (p ((ng there) (vg tense=pres voice=act asp=simple headv=yes...
22.2   (p ((rg up) (pg on) (ng the road) (ng a little bit)))
22.3   (p (and (ng they) (vg tense=presorbase voice=act asp=simple...
22.4   (p (and (vg tense=presorbase voice=act asp=simple headv=yes uh)...
23.1   (p (and (ng they) (vg tense=presorbase voice=act asp=simple...
23.2   (p (and (ng you) (vg tense=presorbase voice=act asp=simple...
23.3   (p ((vg tense=presorbase voice=act asp=simple headv=yes help)...
23.4   (p (and (ng the one thing) (ng that) (vg tense=past voice=act...
24.1   (p ((vg tense=pres voice=act asp=simple headv=yes is) (sg that)...
24.2   (p (but (ng it) (vg tense=pres voice=act asp=simple headv=yes...
24.3   (p (and (ng a ball)))
24.4   (p ((vg tense=pres voice=pass asp=simple is headv=yes attached)...
24.5   (p (and (ng you) (vg tense=presorbase voice=act asp=simple...
24.6   (p (and (ng that sound) (vg tense=past voice=act asp=simple...
25.1   (p ((rg well anyway)))
26.1   (p ((ng so um tsk) (ng all the pears) (vg tense=pres voice=pass...
26.2   (p (and (ng he) (vg tense=pres voice=act asp=simple headv=yes...
26.3   (p (but (ng his hat fe) (vg tense=pres voice=act asp=simple...
```

Phase 2. Feature extraction

Feature extraction is the most important part of the algorithm. In this phase, each prosodic phrase is assigned values for five binary features. The first four features correspond to unique participation frameworks, i.e., configurations of participant roles and relations, which are designed to be representative of common, unique activity types in the Pear Stories. (The configurations and activity types were identified manually, using corpus analysis on a development set of two dialogues.) The first four features are mutually exclusive (only one of the four can be true for any given prosodic phrase). The fifth feature simply indicates past tense, which means it does not represent a complex participation framework, but rather an atomic participant-relational feature. The following explains these five features in more detail.

First-person meta-narrative [1P] This participation framework reflects direct speaker-addressee meta-discussion, e.g., "Did I tell you that?" The intention here is to capture instances where the speaker is no longer talking *about* the movie, but rather when they are engaged in direct meta-talk with the interlocutor.

Generic second-person narrative [2P-GEN] This participation framework reflects a narrative activity told from the perspective of a generic participant, e.g., "You see a man picking pears." This occurred in the development set with some regularity, and it was apparent that episodes of this kind of narrative presentation were strongly associated with intentional segments.

Third-person stative/progressive narrative [3P-STAT] This feature is designed to distinguish narrative activities relating to 'setting the scene', e.g., "[There is a man | a man is] picking pears." This usually occurred at the beginning of a dialogue, and reflects a kind of 'set-up' or preparatory narrative activity.

Third-person event/sequence narrative [3P-EVENT] This feature is designed to distinguish sequential, event-driven third-person narrative activities, e.g., "The man drops the pears." This appears to have been the default narrative activity type in the development set.

Past/non-past [PAST] This is an atomic participant-relational feature intended to distinguish narrative activities in terms of participants' temporal relationship to the narrative content, e.g., "The man drops the pears" vs. "The man dropped the pears." This feature was employed to test whether changes between past tense and historical present were useful in distinguishing narrative activity types.

Automatic feature extraction works as follows. First, prosodic phrases containing a first- or second-person pronoun in a grammatical subject or object relation are identified using the extracted grammatical relations. (Common fillers like *you know*, *I think*, and *I don't know* are ignored.) Of the identified phrases, those with first-person pronouns are marked 1P, while the others are marked 2P-GEN. For the remaining prosodic phrases, those with a matrix clause are identified using the generated sentence parse. Of those identified, if either (1) its head verb is *be* or *have*, (2) it is tagged by TTT2 as having progressive aspect, or (3) the prosodic phrase contains an existential *there*, then it is marked 3P-STAT. The others are marked 3P-EVENT. Finally, if the matrix clause was tagged by TTT2 as past tense, the phrase is marked PAST. In cases where no participant-relational features are identified (i.e., no matrix clause and no pronouns), the prosodic phrase is assigned the same features as the preceding one, effectively indicating a continuation of the previous activity type.

Phase 3. Similarity measurement

The third phase involves comparing the participant-relational feature vector of each prosodic phrase with a feature vector derived from its previous context. This provides a measurement that indicates whether the prosodic phrase reflects a shift to a new segment or the continuation of a current one. Formally, similarity measurement is calculated according to the cosine similarity $\cos(v_i, c_i)$ between the feature vector v_i of each prosodic phrase i and a weighted sum c_i of the feature vectors in the preceding context, such that

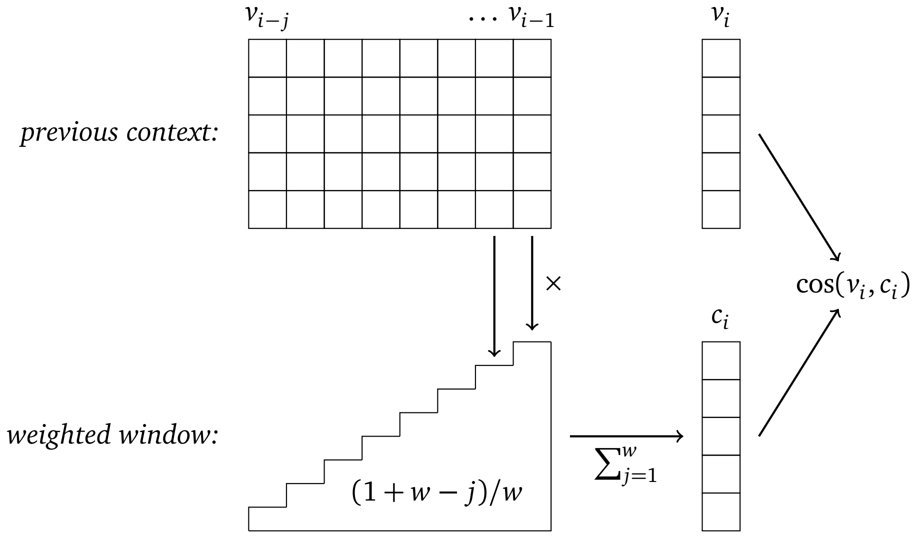
$$c_i = \sum_{j=1}^{w} ((1 + w - j)/w) v_{i-j} \tag{5.1}$$

where a parameter l (representing the desired mean segment length, set by the user) is used to calculate a window length $w = \lfloor (l/2) \rfloor$ of preceding utterances to use as context. The similarity measurement process can be represented graphically as in Figure 5.1. Note how the context window gives increasingly greater weight to more recent phrases.

Phase 4. Boundary assignment

The final phase involves using the similarity scores for the entire dialogue to compute a set of boundaries. To do this, the algorithm simply assigns boundaries to prosodic phrases where $\cos(v_i, c_i)$ is less than the first $1/l$ quantile of the set of such scores for the discourse (i.e., it finds the appropriate number of minimum-similarity boundaries for producing a segmentation with the user-designated mean segment length parameter l).

Figure 5.1: A graphical representation of similarity computation in NM09.



5.2.3 Method

The aim of the experiment is to establish whether participant-relational features are an effective means for intentional segmentation. To establish grounds for what is meant here by "effective," the experiment compares the performance of the proposed NM09 algorithm with a well-known reference-based method (P&L's NP) (in this section also referred to as NP-HUMAN). P&L's NP algorithm works by exploiting human annotations of direct and inferred relations between noun phrases in adjacent units. Inspired by Centering theory (Grosz et al., 1995) and psycholinguistic research (Levy, 1984), these annotations are used in a computational account of discourse focus to measure coherence between utterances.

P&L's approach is based on two main hypotheses. First, they hypothesize that adjacent utterances in the same segment are more likely to contain coreferential or inferentially linked expressions than utterances in different segments. They also hypothesize that speakers are more likely to use definite pronouns than full NPs when referring to an entity mentioned in the current segment (if not mentioned in the immediately previous utterance).

In addition to replicating P&L's original experiment and algorithm, the current experiment also implements a fully-automatic version (NP-AUTO) of the NP-HUMAN algorithm. To create this fully-automatic version, a state-of-the-art coreference resolution system (Poesio and Kabadjov, 2004) is employed to generate the coreference information that the

NP algorithm requires for input. This system achieves between 55% and 77% F-score in an anaphora resolution experiment using the descriptive text portion of the GNOME corpus (Poesio, 2004b).

To provide a top-line for expected performance, results are also calculated for P&L's original human segmentation annotations (HUMAN). To compute this top-line score, the mean of the seven annotators' scores with respect to the three-annotator gold standard is computed. Finally, as a naive baseline, results are calculated for a random assignment of boundaries at the same mean frequency as the gold-standard annotations (RANDOM), i.e., a sequence drawn from the Bernoulli distribution with success probability $p = 0.169$ (this probability determines the value of the target segment length parameter l in our own algorithm).

5.2.4 Evaluation

The experiment follows the suggestions of the evaluation study described in Section 4.2. It applies κ as a chance-corrected performance measure of segmentation boundary identification. Note that this is Cohen's (1960) original agreement measure, not the windowed partial-credit version, a.k.a, k-κ. Rather, in this experiment the evaluation window length k is set to 1, requiring that the placement of segment boundaries be exactly correct for credit to be given. This renders this particular use of k-κ as equivalent to Cohen's κ.

An un-windowed evaluation was chosen because the segments are very short, and it seemed unwarranted to award partial credit in such a fine-grained segmentation experiment. Note, however, that results are also given using standard un-windowed measures F_1, recall, and precision. These are useful for comparing with previous results in the literature and provide a more widely-understood measure of the accuracy of the results. Precision and recall are also helpful in revealing the effects of any classification bias the algorithms may have.

The results are calculated for 18 of the 20 narratives, as manual feature development involved the use of two randomly selected narratives as development data. The one exception is NP-HUMAN, which is evaluated on the 10 narratives for which there are manual coreference annotations.

Table 5.2: Results for intentional segmentation of the 18 test narratives in the Pear Stories.

	κ	F_1	Recall	Precision
HUMAN	.58	.65	.64	.69
NP-HUMAN	.38	.40	.52	.46
NM09	.11	.24	.23	.28
NP-AUTO	.03	.27	.71	.17
RANDOM	.00	.15	.14	.17

5.2.5 Results

Evaluation results for the 18 narratives, calculated in comparison to the three-annotator gold standard, are shown in Table 5.2. The presented values are means over the 18 discourses, and they are ranked according to κ. Significant differences are computed using a paired Student's t-test, with $p \leq 0.05$ the required level for significance. According to this test, the three algorithms (not including HUMAN and RANDOM) are significantly different from each other in terms of κ. NP-HUMAN and NM09 are both significantly better than the random baseline for all measures. NP-AUTO, however, is only significantly better than RANDOM in terms of recall and F_1.

5.2.6 Discussion

The results indicate that the chosen set of participant-relational features are helpful for intentional segmentation. While the results are not near human performance, it is encouraging that such a simple set of easily extractable features achieves results that are 19% (κ) and 18% (F_1) of human performance, relative to the random baseline.

The other notable result is the very high recall score of NP-AUTO, which helps to produce a respectable F_1 score. However, a low κ reveals that when accounting for class bias, this system is actually not far from the performance of a high recall random classifier. Error analysis shows that the reason for the problems with NP-AUTO is the lack of reference chains produced by the automatic coreference system. While the system seems to have performed well for direct coreference, it does not do well with inferred (bridging) relations.

One can conclude from this that inferred relations are an essential part of the reference chains produced by P&L's coreference annotators, and that they play a significant role in the performance of the NP algorithm. One can therefore infer that naive referential relations, such as those based upon chains of identical lexemes, are unlikely to be effective on their own. Rather, a much deeper notion of referential linking is required. Unfortunately, identifying inferred relations automatically is extremely difficult, and requires extensive world knowledge in the form of training on large datasets or the use of large lexical resources. This makes an accurate, fully-automatic implementation of NP a challenge.

The participant-relational algorithm, in contrast, is not dependent on referential links. In fact, it operates over a dimension of the discourse that is largely *orthogonal* to the approach of NP. While it does not perform as well as NP, the results are positive. Certainly, a more robust, domain-independent instantiation will be needed for it to work effectively on meetings and coarse-grained segmentation problems. Still, this initial experiment can be interpreted as a step forward, and provides reason for continuing with further experiments.

5.3 Experiment 2: Topic-based Segmentation of the Pear Stories

The work of Passonneau and Litman (1997) embodies an approach to discourse segmentation motivated by an *intentional* theory of dialogue structure (Grosz and Sidner, 1986). The annotations they produced were created by naive coders employing an informal notion of *speaker intention*. The approach they took toward automatically reproducing that segmentation was motivated by Centering theory (Grosz et al., 1995) and the hypothesis that intentional structure is exhibited in the attentional relationships between discourse referents. One of the main conclusions they found (also supported by our results in Experiment 1) was that inferred relationships between references in adjacent utterances are strongly correlated with segmentation.

P&L's work on intention-based segmentation is, of course, not the only kind of study addressing automatic discourse segmentation. Research in this area has also addressed other types of segmentation problems, such as story segmentation, paragraph segmentation, and topic segmentation. But unlike P&L's work, this other work is not oriented toward the study of intentional segments nor the validation of Grosz and Sidner (1986). To the author's knowledge, P&L is in fact the only work describing an algorithm designed specifically for fine-grained *intentional* segmentation.

5.3.1 Cohesion, intention, and discourse segmentation

One of the most widely-cited topic-oriented approaches to segmentation is the TextTiling algorithm (Hearst, 1997). Designed to segment texts into *multi-paragraph subtopics*, it works by operationalizing the notion of lexical cohesion (Halliday and Hasan, 1976). Many algorithms share this grounding in cohesion. This goes back at least to Morris and Hirst (1991), whose contribution was to show empirically that lexical cohesion is stronger within a discourse segment than between segments. Recent improvements to this method include the use of alternative lexical similarity metrics like LSA (Choi et al., 2001) and alternative segmentation methods like the minimum cut model (Malioutov and Barzilay, 2006) and ranking and clustering (Choi, 2000). Recently, Bayesian approaches that model the words in a segment as generated from global lexical distributions have been employed (Purver et al., 2006b; Eisenstein and Barzilay, 2008). These latest approaches produce robust results that capture far-reaching lexical relationships across several utterances, though they still cite lexical cohesion as a motivating principle.

It is interesting to note that while Hearst and P&L try to achieve different goals, their algorithmic approaches have much in common. Both algorithms employ the notion that segment boundaries should be drawn at locations where links across sentences are minimal. In Hearst's case, links are based upon word repetition. In P&L, the links are based upon coreference, pronouns, and bridging relations. There are, however, clear differences between Hearst and P&L as well. First, there is an obvious difference in the target dataset. Hearst studies written expository text. P&L study spoken narrative monologue. The annotation instructions also reflect an important difference in motivation. Hearst instructed naive annotators to mark paragraph boundaries "where the *topics* seem to change," whereas P&L asked naive annotators to mark prosodic phrases where the speaker had begun a new *communicative task*. The resulting annotations also exhibit a difference in granularity, with intentional segmentation relating to finer-grained structure. Hearst's segments have a mean of about 200 words to P&L's 40.

The subtle relationship between topic and intention, in terms of motivation and algorithmic design, has sometimes resulted in a lack of clarity in the literature. For example, P&L's work has been frequently cited as a study of topic-oriented segmentation, e.g., (Galley et al., 2003; Eisenstein and Barzilay, 2008). Also, recent research in conversational genres (Galley et al., 2003; Hsueh and Moore, 2007b) analyze events like 'discussing an agenda' or 'giving a presentation'. These events resemble more intentional (i.e., purpose or activity-oriented) categories, though they are discussed by the authors in topic-oriented terms. This confusion suggests that further analysis is needed to understand the deeper relationships between different types of coherence and different types of discourse segmentation.

5.3.2 Method

The current experiment is designed to determine whether topic-oriented approaches to segmentation (those that look at the statistical distribution of 'content' words over multiple adjacent utterances) are effective for fine-grained intentional segmentation of the Pear Stories. Relating this to the proposed formulation of participant-relational analysis, such approaches may be seen as focused on *subject matter*. In the participant-relational model, subject matter remains an important component of what might distinguish communicative activities. It can therefore be hypothesized that topics would remain moderately effective on this task, even though the target is an intention-based segmentation. It is useful, however, to test their *relative* performance against the participant-relational approach.

One reason for this is that the participant-relational approach, instantiated here by the NM09 algorithm, may be seen as *orthogonal* to the topic-oriented approach. In fact, the features making up input to NM09 (tense, aspect, and personal pronouns) are explicitly excluded from a topic-oriented analysis. Topic-oriented approaches eliminate *stop words* such as *I* and *you*, and they eliminate tense and aspect through *stemming*. Therefore, an experiment such as this one, which compares performance between these two approaches, is effectively evaluating the relative usefulness of two *independent* sets of features.

The method employed in this experiment is the same as in Experiment 1. The same set of 18 Pear Stories are used. However, instead of comparing the participant-relational approach to P&L's algorithm, it is compared with existing algorithms designed for *topic*-based segmentation. The set of topic segmentation algorithms used here is similar to the set used in the empirical study of segmentation evaluation in Section 4.2.4. However, LCSeg (Galley et al., 2003) is not used due to the extreme bias discovered during previous evaluation, and Hearst's TextTiling algorithm (Hearst, 1997) is added. These algorithms represent a collection of state-of-the-art approaches that may be seen as derivative of the Hearst approach. They perform stop word elimination and stemming in order to arrive at a set of *content word stems* to use as input.

Table 5.3: Segmentation evaluation results for topic-based segmentation of the 18 test narratives in the Pear Stories.

	κ	F_1
NM09	.11 (σ =.09)	.24 (σ =.10)
BayesSeg	.09 (σ =.10)	.24 (σ =.10)
U00	.08 (σ =.13)	.24 (σ =.11)
C99	.08 (σ =.10)	.22 (σ =.09)
TextTiling	.05 (σ =.10)	.18 (σ =.09)
Random	.00 (σ =.12)	.15 (σ =.12)

5.3.3 Results

The evaluation results for the 18 narratives are shown in Table 5.3. All the automatic algorithms except TextTiling are significantly better ($p \leq 0.05$) than the random baseline in terms of κ and F_1. The only statistically significant difference between the automatic methods is between NM09 and TextTiling in terms of F_1. The observed superiority of NM09 over TextTiling in terms of κ was not significant ($p = 0.08$). The observed superiority of NM09 over C99 was also not significant ($p = 0.24$).

5.3.4 Discussion

The comparable performance achieved by NM09 in comparison to lexical-semantic approaches suggests two main points. First, it further validates the participant-relational approach in practical terms, showing that a participant-relational approach can stack up well against a variety of state-of-the-art fully-automatic approaches. Second, the results suggest that topic and intention are related insofar as topics are useful delineators of intentional segments. (Whether the inverse is true—that participant-relational features are useful for topic segmentation—is covered in a subsequent experiment.)

Analysis of the results suggests that the major portion of the Pear Story dialogues are oriented toward the basic narrative activity which is the premise of the Pear Stories experiment. This means that there are many times when the activity *type* does not change at intentional boundaries, i.e., the participation framework (and its instantiation) does not change. Instead, the intentional boundary marks a new activity with a change only in the dimension of subject matter. At other times, however, the activity type (participation framework) does change but the subject matter does not.

Perhaps the most interesting result from the experiments in this section is that algorithms operating over orthogonal dimensions of the dialogue (i.e., topical, referential, and participant-relational), despite doing so, each achieve a significant level of segmentation performance over a baseline. This suggests that they are all useful to some extent, but that none of them are singularly ideal.

Based upon the current analysis, it is, however, difficult to say quantitatively how the different dimensions relate to the various types of segment transitions that occur. In the following experiments, I will study these interactions directly, reaching for a more detailed account of the specific relations between feature type and segment type, and addressing the issue as it arises in workplace meetings.

5.4 Experiment 3: Coarse-grained Segmentation of Meetings

Experiments 1 and 2 showed that a number of distinct feature types can be effective at fine-grained intentional segmentation, including the novel participant-relational features and traditional lexical semantic features. Interestingly, other studies have found an inverse sort of interaction. In an experiment studying *topic* segmentation of meetings (Hsueh, 2008a) (that is, the human-labeled segments were based on topics), using *non-semantic* features performed as well as semantic ones. The use of discourse cue phrases, theoretically predicted to occur at *intentional* segment boundaries (Grosz and Sidner, 1986), have also been demonstrated to have a positive effect on topic segmentation performance (Galley et al., 2003; Eisenstein and Barzilay, 2008).

It is important to begin to tease apart these perplexing associations, and the current experiment proposes to test two hypotheses with this in mind. The first is that two simple participant-relational features—subjectivity and participation—can be used effectively to segment multi-party conversations. The second is that the performance of different feature types will *vary with respect to segment types*.

This type-dependent analysis is simple to perform using the AMI corpus, because it includes annotations of segment labels selected from a small pre-defined set. By identifying which of these labels suggest the occurrence of unique types of participant involvement (e.g., “presentations”), unique semantic topics (e.g., “budget”), or unique subjective attitudes (e.g., “evaluations”), the second hypothesis can be tested.

The experiment is organized as follows. First, a state-of-the-art lexically-driven unsupervised segmentation algorithm is applied to the meeting transcript. In this experiment, I use BayesSeg (Eisenstein and Barzilay, 2008), which showed consistent good performance in the previous evaluation studies. Second, *the same algorithm* is used to segment the meetings based upon *alternative input representations*—each time using input that contains different participant-relational features (the mechanism for doing this is explained below). This divorces the effect of algorithm from the study, allowing us to focus on the effect of input features. Most importantly, the results are then evaluated independently on each of the different segment types (as determined by their labels). The ultimate objective is thus to determine the relationship between different input features and different segment types.

5.4.1 Data

The experiment is conducted using source data from the 138 scenario-based meetings in the AMI corpus (Carletta, 2007). In the complete set of scenario meetings, there are 35 unique subject groups and a total of about 676,000 transcribed words over 64 total hours. Each meeting contains a mean of 848 possible segment boundary locations (boundaries between dialogue acts) and 14 topic segments. (See Section 3.2 for further details about the corpus.)

The experiment compares the performance of three types of features as input to the segmenter: *content word stems*, *subjectivity*, and *participant activity*. Performance is compared against a set of human topic segment annotations. The content word stem and participant activity features are derived directly from the transcript using a simple process that is described in the following section. The subjectivity features and the reference segmentations, on the other hand, are produced from a set of human annotations. In the remainder of this section, I shall describe these annotations.

Topic segment and label annotations

Manually-annotated topic segmentations (Xu et al., 2005) are used as a gold standard reference for the experiment. The segmentations use a hierarchical scheme. For the first level of the hierarchy, annotators are asked to segment the full length of the conversation into contiguous segments according to “what the people in the meeting were talking about.” A loose target of 6–10 segments per meeting is suggested. Annotators may optionally segment the top level recursively into further sub-topic segments, though these are not required to be contiguous nor span the entire parent segment.

Several default labels for the segments are provided in the scheme. These labels are given in Table 5.4. There are three main groups of labels described in the annotator instructions. 'Functional' labels are used to indicate segments where the participants are either discussing the organization of the conversation or discussing something not related to the scenario. 'Top-level' labels represent a set of topics that are expected to occur as a result of the designed scenario. 'Sub-topic' labels are used in sub-topic segments only. The OTHER label is used for segments where none of the other pre-defined labels is appropriate.

It is interesting to take note of some of the properties of this list. One notable property is the inclusion of both activity-oriented labels and topic-oriented labels. By looking for verbs or nominalized activity types, the set of *activity*-focused labels can be identified (shown in underlined bold). By doing this, one observes a trend similar to that found in the analysis of meeting minutes in Chapter 3, where activity-oriented descriptions tend to occur at the larger scales of analysis, and subject matter oriented labels tend to occur at more fine-grained levels.

It is also important to note some incongruencies that were discovered in the annotations. While annotators were asked to use only 'Functional labels' and 'Top Level labels' for segments at the first level of the segmentation hierarchy, they did not strictly follow this guidance. Sub-topic labels occur 38 times as labels for top-level segments (4% of all segments).

Subjective content annotations

The data used as input to the subjectivity-oriented segmenter is derived from twenty meetings in the corpus that have been annotated for subjective content (Wilson, 2008). These annotations mark various types of subjective statements, subjective questions, and objective polar statements. Subjective statements are defined as expressing opinions or sentiments. Subjective questions are defined as *eliciting* opinions or sentiments. Objective polar statements describe positive or negative factual information from which an opinion can be inferred (e.g., "The camera broke the first time I used it"). In total, there are eleven types of subjective content labels. According to Wilson (2008), the annotations have a Cohen's κ inter-annotator reliability rating of between .56 and .68, which may be considered moderately reliable. The annotated statements and questions may cover single words or span sentence-length propositions.

Table 5.4: The pre-defined segment labels and their frequency of occurrence in the corpus. Activity-oriented labels are shown in underlined bold.

Functional Labels		**Count**
F1	**OPENING**	120
F2	**CLOSING**	135
F3	AGENDA / EQUIPMENT ISSUES	195
F4	**CHIT-CHAT**	15
Top Level Labels		
T1	PROJECT SPECS & PARTICIPANT ROLES	51
T2	NEW REQUIREMENTS	32
T3	USER TARGET GROUP	28
T4	INTERFACE SPECIALIST **PRESENTATION**	61
T5	MARKETING EXPERT **PRESENTATION**	61
T6	INDUSTRIAL DESIGNER **PRESENTATION**	59
T7	**PRESENTATION** OF PROTOTYPE(S)	28
T8	**DISCUSSION**	123
T9	**EVALUATION** OF PROTOTYPE(S)	30
T10	**EVALUATION** OF PROJECT PROCESS	25
T11	COSTING	34
T12	DRAWING **EXERCISE**	25
Sub-topic Labels		
S1	PROJECT BUDGET	50
S2	EXISTING PRODUCTS	43
S3	TRENDWATCHING	25
S4	USER REQUIREMENTS	28
S5	COMPONENTS, MATRLS. & ENERGY SRCS.	189
S6	LOOK & USABILITY	255
S7	HOW TO FIND WHEN MISPLACED	20
Other		
O	OTHER	139

With the exception of subjective questions, each subjective content annotation is also marked with two additional pieces of information: the *source* of the subjectivity and the *target* of the subjectivity. Typically, the source of the subjectivity is the speaker, who is expressing his or her own opinion. However, at times the speaker may be speaking for the group or perhaps reporting the opinion of others. The source labels capture these distinctions. The target labels are general categories for indicating what the subjectivity is about. For further details on the subjective content annotations see (Wilson, 2008).

5.4.2 Method

There are three steps in the experiment. In the first step, three separate representations are prepared for each meeting based on the three different feature types: 1) content word stems, 2) participant activity, and 3) subjectivity. For the second step, each feature representation input is segmented using the BayesSeg algorithm (Eisenstein and Barzilay, 2008). Finally, the performance of each feature representation is evaluated for each segment label.

5.4.3 Content words, participant activity, and subjectivity

The input feature representations are formulated as a sequence of lines, with each line corresponding to a single dialogue act (sorted by start time). An example extract of each of the representations is shown in Figure 5.2. This example was chosen to include a segment boundary, which is reflected in observable changes in the underlying features.

The content word stem representation consists of a manual transcript from which stop words have been removed. The remaining words are then stemmed (using the default stemmer and stop word list provided by Eisenstein and Barzilay (2008), reproduced in Appendix C). This representation represents the typical method for lexical-semantic based topic segmentation.

The participant activity representation is designed to reflect the local distribution of participants' verbal contribution to the conversation. This is represented as a single token for each word, where the token contains a sorted, concatenated list of the identities of the speakers of the current and previous utterance. For example, during a presentation by speaker A the token will often be 'AA', whereas 'BC' might occur during a discussion involving speakers B and C. This representation is motivated by the importance of utterance

Figure 5.2: The three feature representations used as input in Experiment 2. [AMI–ES2009c–437]. (*Subjectivity feature key: possubj = "positive subjective", abother = "about other", uncert = "uncertain", abprev = "about previous", qsubj = "subjective question", negsubj = "negative subjective"*)

Original transcript

D: I can pass you on that email from my uh guy in uh guy down the hall .
C: Sounds good .
A: Hmm .
A: What do you think on it Nathan ? About the voice rec ?
B: I think if we we do both the obviously production costs are going to go way up

Content Word Stems	**Participant Activity**	**Subjectivity**
pass email gui gui hall	DD DD DD...	nil nil nil nil nil nil...
sound good	CD CD	possubj speaker abother possubj...
hmm	AC	uncert speaker abprev
think nathan voic rec	AA AA AA...	qsubj qsubj qsubj qsubj qsubj qsubj...
think obvious produc...	AB AB AB...	negsubj speaker abprev negsubj...

pairing, a.k.a., adjacency pairs, when conducting a basic account of verbal contribution, though it rather naively assumes subsequent utterances are paired.

The subjectivity feature representation is a word-level representation of the subjective content annotations. If a word appears in a subjective content annotation, it is represented using the type of subjective annotation, as well as the type of source and target, if present. Words that do not fall within subjective content annotations are given the label *nil*. Use of this representation is motivated by the fact that various types of subjective attitudes (or lack thereof) are likely indicators of the essential nature of a communicative activity.

5.4.4 The BayesSeg algorithm and varying input features

BayesSeg (Eisenstein and Barzilay, 2008) is an unsupervised Bayesian segmentation algorithm that is based on the notion that the words in each topic segment can be modeled as draws from a unique multinomial language model. However, instead of inferring a point estimate for each topic's language model, the algorithm takes a Bayesian approach by employing a Dirichlet compound multinomial distribution and integrating over all possible multinomials. The likelihood of the observed words is defined as a product of the individual likelihoods for each segment, which are formulated to favor compact, consistent distributions, thus aligning the approach with the predictions of lexical cohesion.

The BayesSeg algorithm was initially designed to identify coherent topic segments by virtue of identifying regions with low-entropy distributions of content words. The algorithm, however, need not be applied only to words, as it does not give any consideration to the meaning of the input features, only their distribution. It can therefore be used to find consistent, unique distributions over any multinomial variable. This allows one to apply the algorithm to features such as our participant activity and subjectivity features, which are hypothesized to exhibit regions of low-entropy coherence just as lexical features do.

The algorithm uses a dynamic programming algorithm to find the segmentation which maximizes the observation likelihood, with a quasi-Newton optimization method to infer a single uniform Dirichlet prior for a dataset. For consistency across the different feature types, we do not estimate a prior from the data but employ the suggested default prior of 0.2 across all inputs.

5.4.5 Evaluation

To assess label-dependent performance (i.e., individual segment types), a reference segmentation was created for each label that contains only the boundaries for segments with that label. That is, the entire meeting is re-segmented such that boundaries occur only at the beginning and end of segments of the target type. For all of our evaluations, a hypothesis segmentation is generated that contains the same number of boundaries as the reference.

5.4.6 Results

The results of our experiment are given in Tables 5.5 and 5.6. Each row represents the results for a unique label, and each column represents the use of a unique feature representation. For the evaluation of the participant activity representation, a corpus of 134 meetings is used. (Four meetings were used as a development set.) For comparison with the subjectivity representation, only the 20 meetings with subjectivity annotations are used.[5] Because some topic labels do not occur in every meeting, the results are mean scores for the set of n meetings from the collection in which the label actually occurs. Tests of statistically significant improvements over word-based segmentation were performed using pair-wise Student's t-tests where $p \leq 0.05$. Cases where the participant-relational features

[5] Results for DRAWING EXERCISE are not included because these segments were not labeled for subjectivity.

are an improvement over content word stems are shown in bold (whether significant or not), and the corresponding p-values from the statistical significance test are indicated.

While the goal of this work is to analyze the relationship between individual segment labels and distinct feature types, we also realize the importance of testing whether our participant-relational features are an effective basis for segmentation overall. For this reason, an evaluation was conducted over the full reference segmentation, which includes the boundaries of all annotated segments. For the full set of 134 meetings, k-κ for content word stems was .194 compared to .134 for the participant activity representation. For the subjectivity subset of 20 meetings, k-κ for content word stems was .178 compared to .141 for the subjectivity representation. While neither participant-relational feature is as effective as content word stems, both are substantially better than random, achieving 69% and 79% of the performance of content word stems, relative to random.

5.4.7 Discussion

The results reveal interesting label-feature dependencies. The most noticeable trend is a correlation between the success of the participant activity feature and the labels T4, T5, T6, and T7. These labels all contain the word "presentation." The participant activity feature appears to be successful at identifying the unique participatory structure of presentation activities, where one person is talking most of the time.

Another notable correlation can be seen between the subjectivity feature and labels containing the word "evaluation." One expects that a segment concerning an evaluation activity would contain more prominent usage of subjective language. The results indicate that the subjectivity feature is capable of identifying this property. Subjectivity features also perform well on the two labels USER REQUIREMENTS and LOOK & USABILITY. While these labels do not evoke an orientation toward a unique activity *type*, they do indicate a subject matter which is likely to elicit subjective attitudes and opinions.

As a general conclusion, there is strong evidence in the results for an activity-oriented conception of the nature of *some* discourse segments in meeting conversations. Our non-semantic activity-inspired features perform moderately well, and suggest that employing participant-relational features is a fruitful approach in meetings. The hypothesis that the performance would be highly dependent on the segment type has also been confirmed. This suggests that gold-standard "topic" segmentations of meetings are actually a heterogeneous collection of both semantically- and pragmatically-oriented units.

Table 5.5: Segmentation results (using the evaluation measure k-κ with k=12) for each topic label. The results compare use of the content word stems representation (**Stems**) with use of the participant activity representation (**Participant Activity**), drawing meetings from the full 134 meeting dataset. Values shown are means over the $\boldsymbol{n}$ meetings that contain the topic label. Bold indicates improved performance over content word stems (with p-values shown from a pairwise Student's t-test for significance).

		(full dataset; $n \leq 134$)		
		Stems (k-κ)	**Participant Activity** (k-κ)	$\boldsymbol{n}$
Functional Labels				
F1	opening	–.022	**.012** (p = .005)	118
F2	closing	–.015	–.022	121
F3	agenda / equipment issues	.077	.075	66
F4	chit-chat	–.149	**.006** (p = .111)	14
Top Level Labels				
T1	project specs & participant roles	.098	.015	36
T2	new requirements	.027	.024	25
T3	user target group	–.008	–.024	20
T4	interface specialist **presentation**	.098	**.237** (p = .001)	59
T5	marketing expert **presentation**	.104	**.203** (p = .013)	61
T6	industrial designer **presentation**	.076	**.125** (p = .094)	59
T7	**presentation** of prototype(s)	.055	**.173** (p = .016)	26
T8	discussion	.050	.019	88
T9	evaluation of prototype(s)	.115	**.144** (p = .300)	26
T10	evaluation of project process	.082	–.004	23
T11	costing	.098	.061	30
T12	drawing exercise	.339	.063	25
Sub-topic Labels				
S1	project budget	.208	.025	41
S2	existing products	.029	–.006	33
S3	trendwatching	.056	.040	23
S4	user requirements	.056	.048	25
S5	components, matrls. & energy srcs.	.037	**.053** (p = .219)	78
S6	look & usability	.081	.051	85
S7	how to find when misplaced	–.028	–.022	16
Other				
O	other	.115	.057	39

Table 5.6: Segmentation results (using the evaluation measure k-κ with k=12) for each topic label. The results compare use of content word stems (**Stems**) with use of subjectivity features (**Subjectivity**), drawing meetings from the 20 meetings for with subjectivity annotations. Values shown are averages over the n meetings that contain the topic label. Bold indicates improved performance over content word stems (with p-values shown from a pairwise Student's t-test for significance).

		(subj. subset; $n \leq 18$)		
		Words (k-κ)	**Subjectivity** (k-κ)	n
Functional Labels				
F1	OPENING	–.017	–.018	18
F2	CLOSING	–.021	–.014	18
F3	AGENDA / EQUIPMENT ISSUES	.116	.034	18
F4	CHIT-CHAT	–.007	**.029** (p = .140)	6
Top Level Labels				
T1	PROJECT SPECS & PARTICIPANT ROLES	.039	.029	6
T2	NEW REQUIREMENTS	–.028	–.019	5
T3	USER TARGET GROUP	–.036	–.036	6
T4	INTERFACE SPECIALIST PRESENTATION	.157	.025	10
T5	MARKETING EXPERT PRESENTATION	.016	–.026	10
T6	INDUSTRIAL DESIGNER PRESENTATION	.082	.030	9
T7	PRESENTATION OF PROTOTYPE(S)	–.015	–.015	3
T8	DISCUSSION	.045	.038	15
T9	**EVALUATION** OF PROTOTYPE(S)	–.015	**.225** (p = .105)	3
T10	**EVALUATION** OF PROJECT PROCESS	.051	**.223** (p = .160)	4
T11	COSTING	.121	.082	5
T12	DRAWING EXERCISE		—	
Sub-topic Labels				
S1	PROJECT BUDGET	.176	.041	7
S2	EXISTING PRODUCTS	.017	.011	7
S3	TRENDWATCHING	.019	–.022	2
S4	USER REQUIREMENTS	.107	**.141** (p = .360)	5
S5	COMPONENTS, MATRLS. & ENERGY SRCS.	.118	.012	18
S6	LOOK & USABILITY	.109	**.132** (p = .174)	18
S7	HOW TO FIND WHEN MISPLACED	–.023	–.023	5
Other				
O	OTHER	.167	.127	18

The main conclusion that can be drawn from these results is that the coarse-grained segmentation of conversations likely requires a mixed pragmatic-semantic approach. However, an approach that isolates the various dimensions, with an eye toward the goals of different applications, may be even more appropriate. For example, if one seeks to identify *presentations* within a set of meetings, then participant activity features can and should be used instead of words. Alternatively, if one seeks to identify *debates*, for example, then subjectivity is likely to be better than words. If one seeks to identify both, or to identify potential candidate activity types, then a deeper exploration of the space of participant-relational features will be a fruitful path.

5.5 Experiment 4: More Participant-relational Features

The results of Experiment 3 suggest that the efficacy of various features for segmentation depends critically on the basic characteristics of the segmentation. In this experiment, I continue to explore this relationship between segment type and feature input, with the purpose of exploring a broader range of participant-relational features. The two participant-relational features explored above—subjective language and participant activity—were special in that there were some clear hypotheses that could be drawn *a priori* in terms of their relationship to the AMI segment label set. There were clear examples (namely, those labeled as presentations and evaluations) in which participant activity and subjectivity were likely to play a role. Other participant-relational features, however, (e.g., tense, personal pronouns, and modals) do not have readily apparent connections to the AMI label set. This experiment is therefore an exploratory one, in which the results will be used to identify correlations that might have a sensible interpretation *post facto*.

5.5.1 Method and algorithm

This experiment follows the same method as Experiment 3. However, rather than use BayesSeg (Eisenstein and Barzilay, 2008) as the segmentation algorithm, Utiyama and Isahara (2001) (henceforth UI01) will be used instead. The reason for this is pragmatic. First of all, both segmenters are more accurate than any of the others that have been tested. Both use a similar approach—segments are modeled as multinomial distributions over words, and the optimal segmentation is the one which minimizes the entropy of those distributions (normalizing for segment length). The difference is that BayesSeg draws these multinomials from a Dirichlet prior, whereas Utiyama and Isahara (2001) uses a point estimate of the distribution. The performance effects of this difference appear to slightly favor BayesSeg, considering the results described in Section 4.2. However, the computational complexity of BayesSeg is increased due to the estimation of the Dirichlet prior. Since this experiment and subsequent experiments require extremely large numbers of segmentations to be performed, I make the practical decision to use Utiyama and Isahara (2001) for all remaining experiments.[6]

5.5.2 Tense, modals, and personal pronouns

I introduce three new participant-relational features in this experiment. (An example of the input representations can be seen in Figure 5.3, which is based upon the same extract presented in Figure 5.2 in the previous section.)

The first feature encodes *tense and aspect*. To produce the tense and aspect feature representation, each dialogue act in the raw transcript is part-of-speech tagged (Toutanova et al., 2003). This tagger achieves a 97.2% accuracy on the WSJ section of the Penn Treebank. A representation is then made which includes all and only verb tags. The second feature encodes *personal pronouns*. The personal pronoun representation simply includes any personal pronouns identified in the raw transcript. The third feature encodes *modality*. Its representation contains all raw transcript tokens that are tagged as modals (MD) by the part-of-speech tagger. Finally, *automatically extracted subjectivity* is added as a fourth feature. This feature is produced by running a subjectivity classifier on each dialogue act.[7] A variant of each of these features is also introduced, in which a label identifying the speaker is attached to each input token. These variants are shown in the bottom section of Figure 5.3, and are henceforth referred to as SPKR+.

5.5.3 Evaluation

Two types of evaluation are performed. The first is an "overall" segmentation evaluation, in which each feature is assessed for performance against the original gold standard. In this case, the gold standard is one based upon the highest level of the hierarchical seg-

[6] It should be noted that results from the subsequent experiments, some of which were compared against BayesSeg, actually favor Utiyama and Isahara (2001) in terms of performance. The reason for this is unclear, since BayesSeg is formally a Bayesian generalization of Utiyama and Isahara (2001).

[7] This system was in early development at the time of publication. It was provided by Theresa Wilson and its accuracy on our target data is unknown.

Figure 5.3: Participant-relational features used in Experiment 4. [AMI–ES2009c–437]. This is the same passage as in Figure 5.2. (*Tense and aspect feature key (verb POS tags): VBP = "Verb, non-3rd person singular present", VBZ = "Verb, 3rd person singular present", VBG = "Verb, gerund or present participle"*)

Original transcript

D: I can pass you on that email from my uh guy in uh guy down the hall .
C: Sounds good .
A: Hmm .
A: What do you think on it Nathan ? About the voice rec ?
B: I think if we we do both the obviously production costs are going to go way up

Tense and Aspect	Personal pronouns	Modality	Auto. Subjectivity
VBP	i you my	can	objective
VBZ			subjective
			objective
VBP	you it		subjective
VBP VBP VBP VBG	i we we		subjective
D.VBP	D.i D.you D.my	D.can	D.objective
C.VBZ			C.subjective
			A.objective
A.VBP	A.you A.it		A.subjective
B.VBP B.VBP B.VBP B.VBG	B.i B.we B.we		B.subjective

mentation annotations (Xu et al., 2005). This choice was made so that the evaluation focuses on coarse-grained segmentation. Results are given using the evaluation measures recommended in Section 4.2. The evaluation window parameter is set to a large value ($k = 41$) in order to assure that segmentations are awarded partial credit given the very coarse reference segmentation. As in all previous experiments, the number of segments in the hypothesis is set to match the number of segments in the reference.

The second type of evaluation matches the label-dependent assessment performed in Experiment 3. A reference segmentation is generated for each segment label, such that the segmentation only includes boundaries at the beginning and end of segments with that label. Hypothesis segmentations with the same number of segments are then evaluated against this reference.

5.5.4 Results

'Overall' results (measured over all segments, i.e., the original gold standard) are presented in Table 5.7, ranked by k-κ. The *content word stem* representation is the best performing, matching the results of Experiment 3. Adding the speaker tag (SPKR+) improves performance significantly ($p < 0.05$) for all features (modality, tense, pronouns) except content word stems ($p = 0.19$). Of the participant-relational features that contain no information relating to speaker (not SPKR+ and not participant activity), automatic subjectivity is the most effective, followed by personal pronouns, tense and aspect, and finally, modality. Modality was the only feature that did not significantly outperform a random baseline ($p = 0.08$).

Results are also presented for two label-dependent evaluations that have particularly interesting results—DRAWING EXERCISE and EVALUATION OF PROJECT PROCESS. These results are shown in Table 5.8 and Table 5.9, ranked by k-κ. The values indicate means over the 25 meetings in which the labels occurred (the set of meetings are different for each label—that both labels occurred 25 times is coincidental).

For the DRAWING EXERCISE label, the worst performing three features (modality, tense, and automatic subjectivity) did not perform significantly better than random ($p \leq .05$). All other features performed significantly better than random. Content word stems performed better than all other features. Personal pronouns were the best-performing participant-subjectivity feature.

For the EVALUATION OF PROJECT PROCESS label, only tense and content words stems per-

Table 5.7: Meeting segmentation results comparing participant-relational features and content features.

Feature	k-κ	k-precision	k-recall
Content word stems SPKR+	.249	.468	.580
Content word stems	.227	.456	.505
Participant Activity	.198	.438	.542
Personal pronouns SPKR+	.178	.423	.528
Tense and Aspect SPKR+	.164	.417	.525
Auto. subjectivity	.135	.407	.402
Personal pronouns	.085	.360	.412
Tense and Aspect	.072	.356	.396
Modality SPKR+	.071	.354	.442
Modality	.028	.324	.392

Table 5.8: Results by feature on segmenting DRAWING EXERCISE segments.

Feature	k-κ	k-precision	k-recall
Content word stems	.402	.510	.510
Personal pronouns	.232	.363	.351
Participant Activity	.219	.355	.349
Personal pronouns SPKR+	.218	.353	.349
Content word stems SPKR+	.173	.320	.320
Tense and Aspect SPKR+	.128	.281	.271
Modality SPKR+	.111	.271	.262
Modality	.054	.226	.217
Tense and Aspect	–.012	.168	.149
Auto. subjectivity	–.064	.138	.105

formed significantly better than random. The *tense and aspect* features performed better than all other features, though the improvement over *content word stems* was not significant.

5.5.5 Discussion

The results suggest an interesting mix of conclusions. In terms of label-independent overall performance, none of the novel participant-relational features outperformed either participant activity or content word stems. This indicates that these two features tend to be the most robust and generally applicable. This falls in line with previous results in the literature (Hsueh, 2008a) that indicate participant activity is a very good indicator of coarse

Table 5.9: Results by feature on segmenting EVALUATION OF PROJECT PROCESS segments.

Feature	***k*-*κ***	***k*-precision**	***k*-recall**
Tense and Aspect	.147	.220	.221
Content word stems	.120	.195	.198
Auto. subjectivity	.031	.124	.100
Personal pronouns	.007	.098	.093
Tense and Aspect SPKR+	.005	.091	.091
Personal pronouns SPKR+	–.003	.086	.086
Content word stems SPKR+	–.014	.072	.074
Modality	–.013	.075	.076
Participant Activity	–.055	.038	.038
Modality SPKR+	–.081	.013	.014

segmentation in meetings. And it confirms previous results in all of our experiments and the literature that content word stems are a very useful means of segmentation in a variety of settings.

The outcomes associated with the participant-relational features are, however, generally positive. All of the participant-relational features significantly outperformed a random baseline, with modality as the only exception ($p = 0.08$). This shows that even very simple participant-relational features are indicators (to varying degrees of reliability) of coarse-grained segmentation in meetings. And since these features are typically absent from the content word stem representation, this suggests that they could be used in concert with content word stems to improve the state of the art.

This conjecture is already partially confirmed by the results of using SPKR+ tags, which when added to the content word stem representation produced the best performing system overall, outscoring all previously-published unsupervised approaches. What is interesting to note here is that adding SPKR+ typically improves recall, without affecting precision. In other words, identifying boundaries using participant activity does not cause the system to sacrifice boundaries associated with subject matter. This suggests that combining features conjunctively may be the appropriate solution for feature combination when using Utiyama and Isahara (2001) or related algorithms, e.g., Eisenstein and Barzilay (2008) (though a corresponding increase in sparsity would set a limit for viability).

The two feature-dependent results display interesting properties. The DRAWING EXERCISE segments show a particular affinity for the use of personal pronouns. This appears to be a result of a rather quirky activity that is a required part of the AMI scenario. Namely, the DRAWING EXERCISE label identifies an "icebreaking" activity where each person is asked to go up to the whiteboard and describe their favorite animal. The linguistic outcome of this unique activity is that individuals frequently use first-person pronouns (particularly the word *my*). For example, consider line 44 onward in **(3)** from Section 3.2 (page 53). This is an interesting confirmation of the motivation for participant-relational features. The DRAWING EXERCISE segments involve participants sharing information about themselves in a unique way. In other words, this is a unique activity *type* with a unique *participation framework*, only coincidentally about a unique subject matter, i.e., animals (a likely reason that content words stems also performed very well). In addition, each participant introduces themselves in sequence during this icebreaker, setting up a unique activity structure, which suggests why participant activity and the SPKR+ tags also appear to help for this label.

The EVALUATION OF PROJECT PROCESS displays rather different outcomes. Perhaps surprisingly, tense features are the most useful for this label. One would expect that automatic subjectivity features would perform well on such segments, just as manual subjectivity annotations did in the previous experiment. However, it appears that the automatic subjectivity features are simply not accurate enough to show any significant relationship.

Upon a detailed examination of this segment type, the relationship to tense features does have a reasonable explanation. The EVALUATION OF PROJECT PROCESS label indicates an activity in which participants are discussing their previous experiences (the work they did throughout the day). For this reason, the past tense is used regularly in these segments. Consider the following extract.

(16) [AMI–ES2005d–310]
310 C: Project process, how do we think that went?
311 C: Are we happy?
312 A: Oh.
313 D: Mm.
314 A: Yeah I think we have a a winning product.
315 C: Okay.
316 C: Evaluation.
317 C: Oh we've been writing this up for m months.

(16) [cont. . .]
318 D: I think it went quite smoothly.
319 C: Uh room for creativity, were we happy with that?
320 D: W I think we were very creative.

Clearly, this segment is likely to be distinct in its use of tense, particularly the last few utterances. It is also apparent that the semantic coherence of this segment is tenuous. The subject matter is sparsely defined and abruptly shifting. This might explain why results for content word stems are mediocre. Still, there is an underlying coherence in terms of evaluation and reference to past events. (Note also the use of *bracketing meta-discourse* in line 310, which will be the subject of subsequent experiments on activity type labeling.)

The results of this experiment provide further nuance to the notion of label-dependence established by Experiment 3. It was found that a diverse array of participant-relational features can be effective (at various degrees), but that this effectiveness is highly dependent on the nature of the activity in the segment. Ultimately this produces a compelling story in support of using participant-relations in the development of discourse segmentation systems.

5.5.6 Practical considerations going forward

Discourse segments are not just a component of theories of discourse structure. They serve a use in technological applications, such as summarization, information retrieval, and search. Dividing a long document or audio recording into shorter pieces can provide for more manageable, homogeneous units for an index or summary. This idea is a fundamental motivation for the segmentation experiments just described. To produce an activity-oriented summary, the temporal extent of communicative activities must be identified.

Automatic discourse segmentation can, of course, be usefully applied in a wide range of spoken-language application settings. Fine-grained intentional segmentation (as in Experiments 1 and 2) is potentially useful for deep understanding of written texts, for use in dialogue systems, or in automated evaluation of coherence in educational settings. Coarse-grained segmentation (as in Experiments 3 and 4) can be useful, for example, for indexing news broadcasts (Rosenberg and Hirschberg, 2006), summarizing lectures (Malioutov and Barzilay, 2006), or tracking tutorial dialogues (Olney and Cai, 2005). Consider an example from the domain of lectures. A coarse-grained lecture segmentation can help to identify the basic steps in a lesson, which can then be used to assist students in reviewing them individually.

Across different target domains, a number of complex relationships may be considered to be at play in performing a segmentation. First, there are a variety of target segment types, including topics, intentions, activities, and transactions. This choice can be influenced by factors such as the intended application, motivating theory, or the setting being studied. Consider the fact that standards for establishing a reference segmentation vary considerably across different settings (not just those studied here). The TDT project, for example, segmented news broadcasts by *story*—"topically contiguous segments in a broadcast"—employing an *event*-based notion of topic designed specifically for news (Linguistic Data Consortium, 2002). Their definition of topic is "an event or activity, along with all directly related events and activities," where an event is "a particular thing that happens at a specific time and place," and an activity is "a connected set of events." For lectures, the content of presentation *slides* has been used as a basis for establishing segments (Malioutov and Barzilay, 2006).

Different granularities of segmentation can also be found, from sub-paragraphs to sections for text, and from a few seconds to several minutes for speech. And as investigated in this thesis, segmentation algorithms vary considerably as well, alternatively employing features such as lexical chains, coreference, prosody, and cue phrases (or participant-relations, as proposed here).

The results from experiments described thus far therefore come with significant notes of caution. If there is one thing to take home from these experiments, it is that discourse segmentation is a nuanced, multi-dimensional phenomena with many interacting parts. *There is unlikely to be one singular solution*. And what might appear to be an ideal solution for one domain, setting, genre, or segment type, may prove less effective in another.

It is also clear from the evaluations presented here that wholesale 'overall' evaluations are of limited use. An exploration of the feature space *in conjunction with* the space of activity types is required. But this likely means that a serious consideration of purpose factors is also needed. Earlier in this thesis, I presented an argument in favor of an activity-oriented type of summary for workplace meetings. I then made the argument that it is based on an intentional account of discourse and showed that it generalizes across two rather different corpora. But still, other settings may require dramatically different orientations (though

I believe the activity-oriented participant-relational approach has significant potential to generalize beyond these two corpora).

Clearly, there are many more ideas to explore on the problem of discourse segmentation. But instead of continuing this line of experimentation (which will be the subject of future work), I shall instead turn to label identification. Ultimately, the goal of this thesis is to move toward activity-oriented summarization of conversations, and this means that making segmentation useful requires that it be attached to an activity-oriented labeling process.

5.6 Experiment 5: Inducing Corpus-level Activity Types

5.6.1 Using bracketing meta-discourse to label activity types

Participant-relational analysis proposes that *expressions of participants' relationships to the dialogue* are useful indicators of 'what's going on' (Goffman, 1974) socially in a conversation. Linguistic features such as *participant activity*, *deixis*, and *subjectivity* are therefore hypothesized to be key indicators of context and a principal means for identifying meanings relevant to social purpose in dialogue. The previous experiments generally confirmed this hypothesis (admittedly, with various degrees of success) with respect to the use of such features in *segmentation*.

I shall now present a slightly different set of experiments studying another type of participant-relational expression that has particular relevance to the problem of *labeling* those segments—*meta-discourse*, i.e., participants' talk about talk. Participants often say things like "that's what I mean" or "that might be your opinion." These expressions objectify the dialogue itself, and *explicitly* place participants in relation to their conversation. Meta-discourse can therefore serve a useful role in the creation of summaries that reflect participants' attitudes about the conversation. (This idea was first introduced in Section 1.2.2 and more fully fleshed out in Section 3.5.)[8]

The idea that meta-discourse is relevant to activity-oriented summarization is inspired by the work of Schiffrin (1980). (It is also supported by initial work on the use of 'meta-comments' in extractive summaries by Murray and Renals (2008).) Schiffrin suggests that meta-discourse often occurs as 'organizational brackets.' That is, participants place meta-

[8] Some examples of meta-discourse were provided in **(3)** on page 54 ("introduce yourself") and **(16)** on page 164 ("project process, how do we think that went?").

discourse at the *boundaries* of discourse segments. She identifies two varieties (amongst others)—'initial' brackets that contain cataphoric reference to what is about to come, and 'terminal' brackets that refer anaphorically to previous discourse. (I shall also refer to these as *forward-looking* and *backward-looking* meta-discourse.)

Schiffrin makes the interesting observation that most examples of meta-discourse contain *discourse deixis*, that is, they employ deictic reference to the discourse itself. This is realized in the use of expressions like *this*, *that*, or *here*, which are used to refer to some previous (or future) element of the discourse structure. Consider one of Schiffrin's examples of 'initial' or forward-looking bracketing—"here's the reason." In this example, the speaker is employing deictic cataphoric reference, "here," to refer to what she is about to say. But the use of deixis is only half of the picture here. The example also reveals that what is about to come next is "the reason." In other words, the meta-discourse expresses a subjective conceptualization of the purpose of the subsequent discourse. Upon hearing this, one can infer that what comes next has a 'reason'-oriented rhetorical relation to the previous dialogue. Now consider a (hypothetical) example of 'terminal' or backward-looking meta-discourse that might occur in a workplace meeting—"that was a great discussion." The speaker here describes "that" (some previous discourse) as "a great discussion," which provides potentially useful information (from the participants' perspective) about what sort of activity just occurred.

By studying such meta-discourse, one can build up a kind of folksonomic conceptualization of the purposes of discourse as they are expressed by participants themselves. What is particularly interesting about such a conceptualization is that it seems to exhibit the same orientation toward *intentionality* that was witnessed in the AMI corpus summaries (see Section 3.2). Participants tend to refer to discourse units as *actions*, *activities*, *intentions*, or *purposes* when using meta-discourse. And it is this connection between what participants express in meta-discourse and what people express in summaries that is a fundamental premise for the current experiments.

The experiments to follow test the hypothesis that participants use meta-discourse at the boundaries of discourse segments. But in addition to this basic test of the existence of bracketing meta-discourse, they also test whether meta-discourse may be used to generate *activity type* labels for surrounding segments. That is, they test whether participants speak about the *activities* they are about to perform, or about the *activities* just completed. The hypothesis may therefore be seen as a restatement of the general argument in favor of an activity-oriented analysis of conversation. It suggests that people conceive of conversations principally as purposeful activity, and that this is evidenced by expressions in both summaries and meta-discourse.

The aim here is, of course, not to produce a traditional summary based upon subject matter, i.e., *topical* headlines. It is to identify *activity type* labels—a required element of a activity-oriented summary. While subject matter is likely to play a role in the use of meta-discourse (as it does in natural summaries), I hypothesize that activity-oriented expressions will dominate subject matter. That is, discourse is fundamentally defined by social action rather than topic, and by using meta-discourse to *refer* to discourse, participants are fundamentally referring to *social activity*.

5.6.2 Method and evaluation

In this first labeling experiment, the aim is to arrive at a set of appropriate activity type labels for a given *corpus*. That is, rather than attempt to identify labels for individual discourse segments, the goal is to identify a *global* set of labels, namely the activity-oriented words in the AMI topic segment label set, e.g., *discussion*, *presentation*, *evaluation*, etc., (see bold words in Table 5.4, page 151). In other words, the experiment will succeed if it identifies these words as the most relevant activity type labels. As with Experiments 3 and 4, the AMI corpus scenario meetings are used as source data, with the same four meetings used as development data.

The algorithm is extremely simple and employs what I shall call a **word-based segmentation** approach. First, all of the words in the corpus are identified to produce a corpus vocabulary. (Note that stop words are eliminated from the transcript and words are stemmed prior to analysis. The vocabulary is thus a set of content word stems, though I shall refer to these simply as ‘words’ or ‘stems’ throughout.) Then, for each word in the vocabulary, segmentations of the meetings are produced in which a segment boundary is placed before all utterances (dialogue acts) containing the word. Note that this technique, along with the others used in this thesis, is *unsupervised* (it does not employ models trained from gold-standard annotation).

The word-based segmentation process produces a large collection of segmentations (many thousands) that can be evaluated against a reference segmentation. Those words whose segmentation achieves the best segmentation score can then be considered the best candidates for segment labels, since they have the highest correlation with segment boundaries. By using the chance-corrected k-κ measure, this may be loosely interpreted as the occurrence of a word having high *mutual information* with the occurrence of a segment boundary.[9]

Multiple 'reference' segmentations are used in this experiment. First, the human-annotated gold standard is used as a reference in order to gauge the viability of the approach. As with previous experiments using AMI, segments in the highest level of the human-produced hierarchical segmentation are used (Xu et al., 2005). Clearly, using the human reference prevents the process from being fully automated. Therefore, the word-based segmentations are also tested against an automatically produced 'reference' segmentation. The automatic segmentations used here are the same as those generated in Experiment 4, where the Utiyama and Isahara (2001) algorithm was applied using a variety of participant-relational features.

One would predict that discourse markers such as *okay* and *well* would perform best in this experiment. These words have often been hypothesized (and demonstrated) to contribute to discourse segmentation (Schiffrin, 1985; Galley et al., 2003; Redeker, 2006; Eisenstein and Barzilay, 2008). Common discourse markers, however, are mostly filtered out in advance in this experiment, since the set of "words" used are actually the set of stop-word-filtered word stems (Appendix C contains a list of the words that were removed from analysis). There are other reasons, however, that common discourse markers might not perform well here. Most previous experiments have typically only investigated the contribution of discourse markers when occurring *precisely* at the discourse boundary. In this experiment, evaluation is performed using the k-κ measure with a window setting of $k = 12$. This gives words partial credit if they occur in the *vicinity* of a segment boundary (it is assumed that meta-discourse does not always occur exactly at segment boundaries). Discourse markers are also likely to be more frequent than meta-discourse, since they are derived from a small set of words and signal discourse boundaries at all levels of granularity. One would therefore expect that they might have good recall but low precision on this task.

Note that evaluations in this experiment are only performed for those meetings in which the word occurred, and only those words occurring in at least 60 meetings are considered. Also, when the automatic reference is used, the number of hypothesized seg-

[9] Using a formal measure of mutual information might actually be a more appropriate measure in this case, but the development of such a measure for near-miss tolerant segmentation is reserved for future work.

Table 5.10: Results of word-based segmentation, using human-annotated reference. The 15 best-performing words are shown, ranked by k-κ; n indicates the number of meetings in which the word occurred; k=12.

Word	k-κ	k-precision	k-recall	n
present	0.202	0.412	0.213	120
evalu	0.191	0.360	0.213	61
meet	0.167	0.301	0.223	134
project	0.155	0.330	0.178	118
minut	0.124	0.318	0.144	122
thank	0.124	0.341	0.139	109
start	0.109	0.294	0.126	115
discuss	0.105	0.295	0.124	113
close	0.105	0.372	0.097	86
okai	0.104	0.171	0.705	134
design	0.103	0.204	0.260	135
let	0.097	0.229	0.144	125
decision	0.093	0.295	0.095	79
individu	0.083	0.346	0.078	70
conceptu	0.077	0.356	0.070	61

ments is set to match the number of segments produced by the word-based segmentation. Finally, the results are ranked according to k-κ, and the top-ranking words are selected as the candidate activity type labels.

5.6.3 Results

The results of the experiment when using the human segmentation reference are shown in Table 5.10. Results when using the best-performing automatic segmentation (content word stems SPKR+) as the reference are shown in Table 5.11. Results when using automatic segmentation based upon participant activity as the reference are shown in Table 5.12.

5.6.4 Discussion

When evaluating the word-based segmentations against the human reference, the results appear to strongly confirm the hypothesis. The induced list of words suggests that meta-discourse does occur at segment boundaries, and that participants do tend to refer to communicative activity. The two best-performing word stems are *present* and *evalu*. This result matches the hypothesis precisely. However, the word *discuss* appears lower in the

Table 5.11: Results of word-based segmentation, using Utiyama and Isahara (2001) with speaker-tagged (SPKR+) content word stems as reference. The 15 best-performing words are shown, ranked by k-κ. n indicates the number of meetings in which the word occurred. k=12.

Word	***k*-*κ***	***k*-precision**	***k*-recall**	***n***
thank	0.057	0.101	0.112	108
produc	0.053	0.082	0.089	98
gonna	0.045	0.139	0.157	99
present	0.042	0.092	0.105	119
kai	0.038	0.104	0.112	114
help	0.038	0.065	0.071	62
design	0.035	0.158	0.181	108
draw	0.033	0.091	0.109	79
sell	0.033	0.074	0.092	106
euro	0.032	0.072	0.104	99
budget	0.03	0.060	0.067	62
wanna	0.028	0.081	0.089	100
remot	0.028	0.206	0.226	65
fact	0.028	0.056	0.059	67
price	0.028	0.069	0.078	94

Table 5.12: Results of word-based segmentation, using Utiyama and Isahara (2001) with the speaker activity feature input as reference. The 15 best-performing words are shown, ranked by k-κ. n indicates the number of meetings in which the word occurred. k=12.

Word	***k*-*κ***	***k*-precision**	***k*-recall**	***n***
thank	0.088	0.127	0.143	108
know	0.061	0.219	0.256	78
present	0.055	0.105	0.120	119
kai	0.040	0.105	0.115	114
remot	0.040	0.215	0.274	65
concept	0.040	0.069	0.080	62
try	0.037	0.088	0.093	114
mm	0.036	0.185	0.210	88
thing	0.035	0.196	0.224	97
write	0.033	0.065	0.066	63
technic	0.033	0.069	0.076	94
start	0.033	0.077	0.086	115
tell	0.033	0.060	0.066	89
button	0.032	0.162	0.229	71
requir	0.031	0.071	0.081	110

list than expected, and *exercise* from the label DRAWING EXERCISE is not present in the list.

Analysis of the stem *present* shows that nominalized use, e.g., "I'll show you the presentation" [AMI–ES2004d–40] is more frequent than use as a verb, e.g., "if you want to present your prototype, go ahead" [AMI–ES2002d–19]. The former occurred in the corpus 463 times while the latter occurred 131 times respectively. A similar trend is observed with the stem *evalu*. Nominal use, e.g., "Maybe we should do the design evaluation first" [AMI–ES2002d–524] is more frequent than use as a verb, e.g., "I'll just show you how we're going to evaluate our own feedback to this" [AMI–ES2004d–200]. The former occurred 232 times while the latter 110 times.

One interesting feature to note about these examples is the use of pronominal person reference. Participants refer to themselves and each other within these sentences in a way that reflects their participation in the activity. These words could therefore also be useful in establishing a role structure to be used in the activity description. Also, the frequent use of nominalizations suggests that using stems rather than verbs was an appropriate choice, but that interpretation requires recognition of light verb or phrasal verb constructions.

It can be reasonably hypothesized that the stems *start* and *close* might have something to do with the meeting OPENING and CLOSING activity labels. However, analysis shows that the *start* stem plays a role in more generic forward-looking bracketing in phrases such as "you wanna start us off?" [AMI–ES2010a–138]. In a similar manner, the stem *let* is also not clearly associated with any particular activity type. Rather, it too is a generic forward-looking bracketing device when used as *let's*, as in the suggestion for future action "let's start from the beginning" [AMI–ES2006b–225].

The induced list of words also contains other interesting activity- or action-oriented words not appearing in the human label set, such as *thank* and *decision*. Transcript analysis shows that the phrase "thank you" occurs frequently at the end of meetings, the end of presentations, and the end of introductions during the icebreaker. This affirms its use as a discourse bracketing device (and potential use in segmentation), but it is not a particularly informative expression in terms of characterizing the activity being referred to. Analysis of *decision* shows it is used as both forward-looking and backward-looking meta-discourse. One particularly interesting example of this was "but we need to make a decision about um the things we've discussed" [AMI–ES2004c–688]. This refers back to previous discourse as the subject of the decision, but forward toward the discourse that will be required to make the decision.

The word list is also populated with some labels associated with subject matter. The stems *project*, *individu* and *conceptu* seem to indicate regular themes from the AMI scenario, though it is unclear why some of these correlate with segment boundaries—they do not match with the subject-matter labels in the human label set. One observation is that *project* is used frequently at the beginning of the set of four meetings in the scenario, e.g., “well this is the kick-off meeting for our project” [AMI–ES2002a–3]. This example also suggests a reason for the stem *meet* being included in the list as well.

The stem *minut* is a particularly interesting case. It appears to be associated with time-keeping usage at the beginning and ends of the meetings, e.g., “we’re gonna have to wrap up pretty quickly in the next couple of minutes,” but also with programmatic activities, such as reading minutes from previous meetings or talking about their production at the end of a meeting, e.g., “I’ll just recap on the minutes of the last meeting.” In this case, ‘recap of minutes’ might be an appropriate activity label, but in the others, the use of *minutes* is purely correlative and not all that informative of the surrounding segment’s purpose.

Note also that the stems *okai* and *kai* indicate use of the words “OK” and it’s shortened form “kay”, respectively, whose presence in the list confirms the hypothesis that some general discourse markers would be identified by the approach. It is interesting to note its rather distinct recall and precision scores (low precision, high recall). No other discourse markers are present in the list, likely due to their inclusion in the stop word list.

Evaluation against the automatically-produced reference

The word lists generated by use of an automatic reference are, unfortunately, quite problematic (Tables 5.12 and 5.11). Overall, the k-κ scores are quite low, and the words *thank*, *present*, and *start* are the only activity-oriented words matching the list generated from using the human reference. This suggests that the automatic segmentation algorithms are simply not accurate enough to be used on this dataset to produce reliable labels completely automatically. The combination of noise in the segmentation output and noise inherent in the relationship between meta-discourse and discourse segments dominates the information that might be present in the result. This means that moving forward with this approach, as a fully-automated solution, is still a ways off.

However, the positive results using the human reference suggest avenues for progress on automation. For example, the simple fact that particular words do correlate (though very weakly) with automatically produced segmentations, suggests that such an evaluation could be used to gauge the performance of various input features on corpora for which there are no annotations. In other words, if a feature used as input to a segmentation algorithm produces higher correlation to word-based segmentation overall, then it is more likely to be producing a valid segmentation in the first place. This notion also translates to a more general take-home message pertaining to algorithms that model boundary cues explicitly—that one should include bracketing meta-discourse as a boundary-indicative feature in such algorithms. Eisenstein and Barzilay (2008) actually proposes an unsupervised solution of this kind—one which models both intra-segment coherence and term-boundary associations. However, their model forces boundary cues to be present *precisely* at the boundary. Changes to this model that introduce flexibility in the position of boundary cues, or that introduce deeper linguistic features indicating meta-discourse, such as the use of "let's" or personal pronouns, would likely be a fruitful next step.

Clearly, this experiment has not identified a viable end-to-end solution. Still, the basic premise that bracketing meta-discourse exists at segment boundaries (and is activity-oriented) has been confirmed. And at a global level at least, it appears that given an adequate segmentation, one can indeed automatically identify the socially-constituted types of interaction in a corpus. This is a potentially powerful result that could inform approaches to document structure induction, genre identification, and many other problems involving induction of commonsense labels for the *functional* properties of language.

5.7 Experiment 6: Labeling Individual Activities

The lack of positive results in Experiment 5 when using the fully-automated evaluation scheme suggests that further investigation would not be immediately informative unless major changes to the model are made. However, this is not within the scope of the current thesis, as the focus is on discovering discourse patterns rather than automating more robust algorithms. Therefore, in the following final experiment in this thesis, human segmentations are used to ascertain whether bracketing meta-discourse might be used to label activity segments. The aim here is to look at each activity type specifically, and to determine the accuracy of using meta-discourse to label instances of each of them. The hypothesis being tested is that presentation activities will be bracketed by meta-discourse using the stem *present*, that EVALUATION OF PROTOTYPE segments will be bracketed by the stem *evalu*, and so on, and that given a segmentation, k-κ evaluation can be used to discover the label (i.e., word-based segmentation) that correlates best with the boundaries of each segment.

5.7.1 Data, Method, and Evaluation

The experiment requires a segmentation of the corpus that is assumed to be provided. In this case, the human reference segmentation is used. In the experiment, each segment in each meeting is considered individually as a labeling task. For each segment, a new segmentation is created which contains only those boundaries at the beginning and end of the target segment. Then, a collection of word-based segmentations are evaluated against this, and the best-performing word is chosen as the label for the segment. Labeling accuracy is then assessed as a percentage correct.

To establish ground-truth labels for each segment, the activity-oriented words in the human topic-segment annotations are used. That is, each segment is assigned the label *present* (T4–T7), *evalu* (T9 and T10), *discuss* (T8), or *exercise* (T12) (refer to Table 5.4 for the list of segment labels). Note that the experiment does not consider subject-matter oriented labels. Special consideration, however, is given to the 'Functional Labels'. The experiment considers *agenda* to be an activity-oriented label for F3, and the words *start* and *close* are used as targets for the labels F1 and F2 respectively.

The experiment is conducted under two conditions. In the 'open' condition, word-based segmentations based upon each and every word in the vocabulary are evaluated against each segment. In the 'closed' scheme, the automated labeler chooses from a small set of activity-oriented labels that have been hand-selected from the set of labels induced in Experiment 5 (see Table 5.10, page 171). The selected 'closed' class labels are *start, minut, close, agenda, discuss, present, decision, meet, and evalu*. These are chosen by considering the top 15 ranked labels from Experiment 5, and then eliminating light verbs, function words, and topic-oriented words. The 'closed' scheme, because it requires human intervention in selecting the possible set of labels, may be considered a lightly supervised technique, though the supervision task is rather trivial. Note that the experiment only employs human label annotations for evaluation, and is not trained from the set of labels in the topic annotation scheme. A final summary evaluation is also performed that proportionally weights the label-dependent scores by the total number of occurrences of each in the corpus.

Table 5.13: Results of activity segment type labeling, based upon a human segmentation. The table shows, for each segment label type, the percentage correct in the 'open' and 'closed' labeling tasks. A summary evaluation score overall all activity-oriented segments is also provided.

		Open	Closed
Functional Labels			
F1	**OPENING**	16.1	23.7
F2	**CLOSING**	14.0	24.0
F3	**AGENDA** / EQUIPMENT ISSUES	9.0	15.1
F4	**CHIT-CHAT**	0.0	0.0
Top Level Labels			
T4	INTERFACE SPECIALIST **PRESENTATION**	10.7	39.0
T5	MARKETING EXPERT **PRESENTATION**	13.1	45.9
T6	INDUSTRIAL DESIGNER **PRESENTATION**	10.2	42.4
T7	**PRESENTATION** OF PROTOTYPE(S)	23.1	34.6
T8	**DISCUSSION**	10.2	18.1
T9	**EVALUATION** OF PROTOTYPE(S)	42.3	69.2
T10	**EVALUATION** OF PROJECT PROCESS	4.3	47.8
T12	DRAWING **EXERCISE**	0.0	0.0
	ALL ACTIVITY SEGMENTS	**12.4**	**26.4**

5.7.2 Results and Discussion

Results of Experiment 6 are shown in Table 5.13. The results show that for all activity-oriented segments in the AMI corpus (those labelled with *discussion*, *presentation*, etc.), 12.4% of them can be precisely labeled with the correct activity-oriented label, assuming no human supervision at all. If light human supervision is employed to select a smaller number of candidate labels, the accuracy increases to 26.4%. And when considering the more readily discernible types of activities, i.e., presentations and evaluations only, overall accuracy increases to 45.2%.

These results appear rather weak. It is not particularly encouraging that only one quarter of the segments could be correctly labeled in the 'closed' evaluation. However, these results should be considered in light of the complexity of the problem. Participants most certainly do not *always* use meta-discourse to describe what they have done or will do. And so it is not surprising that a technique which relies exclusively on meta-discourse might fail to identify labels for a large number of segments. For example, in introducing another speaker for the INDUSTRIAL DESIGNER PRESENTATION activity, one speaker uttered “very good, and uh, now with David...” [AMI–ES2002c–880.07s]. This is clearly an example of discourse bracketing, but there is no explicit mention of the activity type. (Note interestingly that there *is* mention of the participant!) The technique employed here has no way of harnessing the “now with” discourse indicator for the purpose of labeling.

It should be noted also that the approach is designed as an *unsupervised* summarization technique, i.e., it does not employ the traditional label-and-train approach (though in this case it is leveraging human transcripts and segmentations). This suggests that the approach is extremely generic, and could potentially be used to, for example, identify section headings in document archives. What’s also interesting to note is that the state-of-the-art in *supervised* conversational activity classification, represented by Ries (2001b), achieves very little improvement over a random baseline (p. 142). Clearly, the task is a rather difficult one.

Another possible reason for poor performance is the inherent difficulty of the segmentation problem for humans. (The experiment relies upon human labels.) Creators of segmentation annotations tend not to report inter-annotator reliability, and when they do, agreement scores are often low. Galley et al. (2003), for example, employed a majority vote method of reference segment generation and note that approximately one fourth of the conversations in their corpus did not achieve a reliable level of agreement for inclusion. In addition, their use of Cochran’s Q and a windowed analysis to evaluate reliability is a particularly liberal methodology. The annotations of lectures created by Malioutov and Barzilay (2006) employed an annotation scheme based on Gruenstein et al. (2005). They report agreement in terms of P_k, with scores between .219 and .418, highlighting a large difference in the number of segments assigned by each annotator. They suggest that the highest level of agreement achieved serves as a benchmark for comparison of automatic results. With these sorts of mediocre agreement results, then, it is possible that coarse-grained discourse segmentation is simply a very difficult problem, even for humans. This would have a knock-on effect in the accuracy of any labeling method built upon segmentation, such as the one produced here.

While results in this experiment were not ideal, it is by no means the end of the road in terms of using meta-discourse. One particularly important way forward would be to establish *corpus-level associations* between induced labels and *the features of the segments they bracket*. The results of Experiment 5, in which a set of labels were induced for an entire corpus, were rather encouraging. If each label could be reliably identified as bracketing coherent distributions of (preferably participant-relational) features, then that label could be used to label segments with such a distribution, but around which no such bracketing meta-discourse occurred. This is an obvious next step for moving forward with this technique, though unfortunately, it is not within the scope of the current thesis.

5.8 Summary of Experimental Results

The six experiments described in this chapter provide a consistent and informative set of results concerning the segmentation and labeling of activity-based discourse segments, their relationship to participant-relational features, and their relationship to bracketing meta-discourse.

In Experiment 1, it was found that a simple set of corpus-specific participant-relational features (involving the use of personal pronouns and tense) are useful for the task of fine-grained intentional segmentation. While the results were not near human-level performance, it is encouraging that the very simple NM09 algorithm produced results at approximately 19% of human performance, relative to a random baseline. It was also revealing to find that an automatic version of the NP algorithm (Passonneau and Litman, 1997) was not better than NM09. The participant-relational approach appeared to be more robust than NP, which appeared to depend considerably on inferred (bridging) reference relations.

In Experiment 2, NM09 achieved performance levels comparable to state-of-the-art topic-based (i.e., lexical-semantic) approaches. This further validated the participant-relational approach in practical terms, and revealed that topic and intention have a tendency for confusion (insofar as topics can be useful delineators of intentional segments). An interesting conclusion arising from deeper study of this confusability was that (at least in the Pear Stories) there appeared to be times when the activity *type* did not change at intentional boundaries. Rather, the intentional boundary marked a new activity, but with a change only in the dimension of subject matter, not activity type. At other times, however, the activity type (or participation framework) did change but the subject matter did not. This suggested that the two feature sets may be seen as operating over orthogonal dimensions, suggesting that they are both useful to some extent, but that neither of them are singularly ideal.

Experiments 3 and 4 dove into this orthogonality problem head on, studying the *label-*

dependent effects on the performance of various participant-relational features, and applying the problem to coarse-grained segmentation of workplace meetings (our ultimate objective). The overall performance of participant-relational features was positive. All of the participant-relational features significantly outperformed a random baseline, with modality as the only exception. The most noticeable label-dependent trend was a correlation between the participant activity feature and segments labeled with the word "presentation." Subjectivity features were also shown to correlate (though less reliably) with segments with the label "evaluation." Even tense and personal pronouns were moderately predictive of some isolated segment types, such as DRAWING EXERCISE and EVALUATION OF PROJECT PROCESS. All of this evidence supports an activity-oriented conception of the nature of *some* discourse segments in meeting conversations. It also supports the conclusion that segmentation of conversations likely requires a multi-dimensional, mixed pragmatic/semantic approach.

There was also some preliminary evidence that combining features across these dimensions can improve on the state-of-the-art in unsupervised segmentation. Adding SPKR+ tags to the content word stem representation produced the best performing system overall, outscoring all previously-published unsupervised approaches with a k-κ score of .249. Notable in this result was that adding SPKR+ improved recall, without affecting precision, suggesting that combining features conjunctively may be an appropriate feature-combination approach (when using the entropy-minimizing distributional segmentation approches represented by Eisenstein and Barzilay (2008) and Utiyama and Isahara (2001)).

Experiment 5 then showed that one can induce a list of activity-oriented segment labels using meta-discourse. It showed that meta-discourse does occur at segment boundaries, and that participants do tend to refer to communicative activities. The two best-performing word stems were *present* and *evalu*, which matched the hypothesis precisely. Of course, the induced list also included words like *okay*, which confirms that some general discourse markers are identified by the boundary-correlation approach (even though most were eliminated by the use of a stop word list). The word lists generated by use of an automatic reference segmentation were, however, quite problematic. Moving toward a fully-automated approach therefore requires improvement of segmentation techniques as well. Still, the basic premise that bracketing meta-discourse exists at segment boundaries (and is activity-oriented) was confirmed. It appears that given an adequate segmentation, meta-discourse is a likely feasible way to automatically identify the socially-constituted types of interaction in a corpus. This is a potentially powerful result that could inform approaches to document structure induction, genre identification, and other problems involving learning commonsense labels for the *functional* aspect of language use. One other important conclusion from this experiment was that the windowed partial-credit k-κ evaluation did an effective job as a near-miss tolerant approximation of mutual information between words and segment boundaries.

In Experiment 6, in which labels were identified for individual segments, the results were weaker. The results showed that 12.4% of the segments can be precisely labeled with the correct activity-oriented label, assuming no human supervision at all. When light human supervision was used, the accuracy increased to 26.4%. These results should, of couse, be considered in light of the complexity of the problem and the approach taken. This was an *unsupervised* summarization technique that did not employ the traditional label-and-train approach. It also did not study the use of corpus-level associations between induced labels and the features of the segments they bracket. This is a likely avenue for further progress with this technique.

Bringing all of these results together, particularly those coming from Experiment 4, one can begin to see how different types of features in conversational discourse relate to distinct components of activity descriptions. In Section 3.3, the PAS (i.e., participation, activity type, subject matter) structure was introduced as a general template-like form for sentential activity descriptions. This was then supplemented with a specific analytical method involving participant-relational features (summarized most succinctly in Figure 3.6). Having tested our analytical method against ground-truth activity descriptions, we can now witness this information pipeline from beginning to end—from participants' expressions of their relationships to the dialogue to natural language activity summaries. Our analysis of this pipeline, and the relationships between inputs and outputs that it has helped to reveal, is summarized in Table 5.14. In the left column of the table are the types of linguistic features that have been investigated throughout the experiments. In the right three columns, each of these input features is associated with one or more of the three components of the PAS structure, and their potential for informing the resulting activity description component is provided. Information presented in the table is for the most part concretely established by the experimental results, though some cells in the table (marked with an *) are suggested but less concretely established.

As the table shows, the majority of our conclusions relate to identifying or distinguish-

Table 5.14: Summary of relationships between discourse features and activity description components.

Discourse feature	Participation	Activity type	Subject matter
Subjectivity	*For evaluations, determines who is evaluating	Identifies evaluations, distinguishes informal speech (i.e., CHIT-CHAT in AMI corpus)	Identifies subjective topics, e.g., LOOK & USABILITY in AMI corpus
Modality	—	A mildly informative supporting feature.	—
Level of activity	Distinguishes active participants from inactive ones.	Identifies presentations, or other dialogues involving only some of the participants.	—
Pronouns & reference	*Identifies primary participants or groups of participants	Distinguishes self- and other-focused activity types, e.g., narrative modes in Pear Stories.	—
Content words	—	—	Essential feature for subject matter extraction.
Tense and aspect	—	Distinguishes between past, present, and future-oriented activities, e.g., narrative modes in Pear Stories.	

ing the activity type component of the PAS structure. This was by design—the experiments were designed specifically to illuminate the relationship of the activity type PAS component to the source data. Experiments 5 and 6 on meta-discourse bracketing, though also designed with activity type identification in mind, ultimately found that subject-matter can also be summarized using the same technique. This suggests a more generic usefulness of meta-discourse bracketing to summarization in general. Indeed, this conclusion matches well with findings from work on argumentative and rhetorical summarization of scientific literature, where document sections like *results* or *conclusions* are identifiable often because they are prefaced or concluded by indicative phrases like “we aim at,” or “we have shown that” (Teufel and Moens, 2002).

Chapter 6

Summary of Contributions and Future Work

6.1 Summary of Contributions

This thesis presented a qualitative and quantitative inquiry into the *activity-oriented* nature and structure of spoken conversation.

My main theoretical contributions were presented in Chapter 3. In that chapter, I argued that conversations are composed of coarse-grained episodes of socially-constituted activities. I showed that such activities are important because they are part of participants' commonsense understanding of what happens in a conversation. I showed how they appear in natural summaries of conversations such as meeting minutes, and that participants talk about them within the conversation itself. I also developed a prototypical semantic framework for activity descriptions called PAS. The PAS structure consists of a participation component, an activity type component, and a subject matter component. I showed how this structure relates to verb semantics and role structure, and I suggested possible ways in which such a template-based approach might constrain the difficult problem of discourse-oriented summarization.

In the final section of the chapter (Section 3.6), I presented a novel analytical framework called *participant relational analysis*. Through qualitative corpus analysis and an interdisciplinary synthesis of prior theoretical work, I argued that communicative activities are principally indicated through participant-relational features, i.e., expressions of relationships between participants and the dialogue. Participant-relational features, such as subjective language, participant reference, and participants' speech activity, are therefore argued to be a principal means for analyzing the nature and structure of communicative activities.

In Chapter 4, I described two studies that were important pre-requisites to the main experiments. Section 4.1 presented a study of manual annotation of participant reference. The novel scheme proposed new distinctions for vagueness, discourse function, and addressing-based referent inclusion, allowing a finer-grained analysis that is hypothesized to reflect participation frameworks in a more nuanced way than traditional annotation schemes. The study showed that these distinctions can be reliably coded, and also provided some insight into the complexity of participant reference resolution problems. The produced dataset included annotations of 11,000 occasions of person-referring, which should prove useful to experimentation on person reference resolution and participant proper name induction.

In Section 4.2, I showed by analytical and empirical means how the commonly-used segmentation evaluation measures P_k and WindowDiff fail to penalize substantially defective segmentations due to inherent biases. I also showed how their definitions are ambiguous and have lead to various interpretations being used in the literature. I therefore proposed a novel chance-corrected evaluation measure k-κ, which was used in all subsequent experiments. Additionally, a re-evaluation of state-of-the-art segmentation algorithms using the novel measure produced substantially different results from previous studies. This raised serious questions about the effectiveness of some state-of-the-art algorithms and the validity of previous experiments.

In Chapter 5, I applied the participant-relational framework to two computational problems: *automatic discourse segmentation* and *automatic discourse segment labeling*. Experiments 1 through 4 tested whether participant-relational features were effective at automatically segmenting conversations into discourse segments, e.g., activity episodes. Results showed that they are effective across different levels of segmentation and different corpora, and indeed sometimes more effective than the commonly-used method of using semantic links between content words, i.e., lexical cohesion. They also showed that feature performance is highly dependent on segment type, suggesting that human-annotated “topic segments” are in fact a multi-dimensional, heterogeneous collection of topic and activity-oriented units.

Experiments 5 and 6 tested whether meta-discourse can be used to automatically identify labels for discourse segments. In contrast to assigning semantic topic labels, such as topical headlines, the proposed “word-based” segmentation algorithm automatically labels segments according to activity type, e.g., presentation, discussion, and evaluation. The method is unsupervised and does not learn from annotated ground truth labels. Rather, it induces the labels through loose correlations (using the k-κ measure) between discourse segment boundaries and the occurrence of bracketing meta-discourse, i.e., occasions when the participants talk explicitly about what has just occurred or what is about to occur. Results show that bracketing meta-discourse is an effective basis for identifying some labels automatically, but that its use is limited if global correlations to segment features are not modeled.

6.2 Future Work

The work in this thesis has been successful at illuminating the importance of activity and multi-dimensionality in discourse structure analysis. It has also provided a fruitful new focus on participant-oriented factors. On the other hand, the work leaves a lot of room for further investigation. That further work might concern participant-relational features or meta-discourse specifically, or it could also be a broader investigation of multi-dimensionality in discourse structure in general. There are three main avenues of research that I see as excellent candidates for pursuit.

6.2.1 From Speech Acts to Genres: Inducing Functional Categories

In many ways, this work may be seen generically as the application of an intentional approach to discourse unit labeling. And for this reason, it has a direct connection (downward) to *speech acts* and (upward) to *genres*. Both of these notions are rooted in an intentional (or perhaps simply functional) consideration of language, but at different scales.

Considering such a generalization of scale, the question arises whether the problem of segmentation is all that important. The fact that so much attention was paid to segmentation here might be considered a mere by-product of the fact that I have studied workplace meetings. Meetings tend to be about an hour long, and under such conditions, people typically try to sequentially accomplish multiple goals. The genre is inherently defined by long concatenated sequences. In this sense, this work might also be comparable then to analysis of textbooks, and the segmentation techniques employed here might also be applied to archives of such extended documents. For example, one will likely find that the occurrence of the label “abstract” in a collection of published academic articles can serve as “meta-discourse bracketing” for segments of a particular identifiable nature. That identifiable nature is likely to include features like participant-relational features—things that reflect the author's stance toward the information. Segmentation in this scenario would be very important.

But in applying this work to genre analysis, segmentation is less important. One might instead look for "non-bracketing" meta-discourse throughout a document, without much consideration for its position. Can one, for example, automatically arrive at a description of the academic journal genre by looking for speech act verbs? What would happen if one were to tokenize, parse, and extract all the verbs in journal articles for which the subject was "we" or "I," and then rank them in a tf/idf style in comparison to their use in a collection of different genres. I hypothesize that this might be a good description of what's "going on" in academic journals. Teufel and Moens (2002) suggest such constructions are at least important for large-scale rhetorical analysis.

The same might be true of speech acts. Researchers have been struggling for years to come up with effective solutions for recognizing speech acts automatically. The inference problem, however, is just too challenging, and the supervised machine learning approaches that are invariably used are just too limited by sparse training data (not to mention by their own imposed taxonomy of acts). Why not, then, use a Hearst-style approach (Hearst, 1998) to search enormous archives of text for explicit patterns, in this case speech act patterns like "I agree. . . " or "We claim that. . . ". Might a learned pattern of local discourse features be indicative of the meaning of those expressions, in the same way that a distributional meaning of lexemes can be learned from local contexts?

6.2.2 Coclustering of Features and Contexts

This consideration leads directly to a generic machine learning model that might be applicable to the set of problems considered in this thesis: co-clustering (Dhillon et al., 2003). The generic outline of the work in this thesis is to associate contextual linguistic features with other local descriptive features that might be used to generate a summary of some block of language (here, it was specifically participant-relational features and meta-discourse). The problem clearly in evidence, however, is that some features are active in some cases and others are active in others. Coclustering provides a model in which features along two dimensions are simultaneously clustered. This can be done in such a way as to optimize the mutual information between the two dimensions. This can therefore be used to simultaneously perform three important tasks: (1) cluster (say, participant-relational) features into mutually informative sets, (2) cluster (say, verbs in meta-discourse) into mutually informative sets, and (3) associate each of these clusterings with each other in a mutually informative way. This might provide an elegant mechanism (lacking in the experiments above) by which to associate feature distributions (such as distributions of personal pronouns or tense) with co-occurring descriptors (such as speech act verbs describing the discourse segment). This is perhaps the lowest-hanging fruit for making Experiments 5 and 6 more robust.

6.2.3 Re-introducing sequence and hierarchy

There are two glaring simplifications of discourse structure that I employed in the work in this thesis. The first is that sequencing within a segment was ignored. The second is that the hierarchical nature of segment structure is also ignored. Both of these are central parts (though perhaps to varying degrees of formality) in most discourse theories. Smith (2003), for example, suggests that the main 'discourse modes' are distinguishable not just by the occurrence of tense, aspect, etc., but also by the way in which sentences are sequenced. This is bread and butter to most discourse theorists, but integrating sequentiality into a global segmentation model is an unsolved challenge. Yes, it might be possible to use a pipelined approach. One could segment a discourse, and then address segment-internal sequencing independently for each segment. This approach, however, does not allow for the global use of sequencing features. The same issue can be seen in the use of a hierarchical model of segmentation. This is perhaps even more challenging to the discourse analysis problem because there is a rather significant debate about whether discourse even has a hierarchical component. It surely does in some places, but in genres like conversational speech, that is much less apparent. These problems are therefore ones which might be addressed in the longer term, most likely by complex statistical-relational models (currently at the forefront of machine learning research).

6.2.4 Ongoing work

Some of this suggested future work, such as coclustering and the use of Hearst patterns is ongoing. The most dramatically new aspect of this current work, however, is that it is being conducted on written texts. The fact is that despite being a rather comprehensive and relatively large collection of spoken conversations, the AMI corpus is in many ways simply too small for this kind of work. In the natural world, we learn to categorize our communicative activities into socially-relevant types through repeated use over many years. It is unlikely that a computer will at any point in the near future be able to learn to reliably identify and distinguish activities like “debates,” “discussions,” and “evaluations” from just a few examples.

Appendix A

Person Reference Annotation Scheme

A.1 Introduction

This document describes a procedure for annotating references to people in conversations. It is intended primarily as an instruction manual for annotators but also serves as a reference for annotation consumers. The goal of the annotation task is to provide gold-standard empirical data about the way people refer to people in conversation, with a focus on how participants refer to themselves and other participants. The resulting annotations will be used by researchers to investigate and model this behavior.

The annotation procedure involves listening to (and possibly watching) a recording of a conversation, proceeding one utterance at a time from start to finish. A transcript is also displayed. Your goal is to identify any occasions when the speaker refers to a person (called a person-referring event). Person-referring events have two main components: the person being referred to (called the person referent) and the words used to do the referring (called the person-referring expression). The following is a summary of the steps for annotating each utterance:

✓ Conceptually analyze the utterance. (Section A.2)

> To identify the person-referring events and their components, your first objective is to perform a conceptual analysis of the utterance. This does not involve any actions in the annotation tool. Rather, it informs the actions you will take in the next steps.

✓ Create referent objects representing any new person referents. (Section A.3)

> The first annotation action requires you to create a referent object representing each real person who is referred to specifically and unambiguously in the utterance and who has not been previously referred to in the conversation. You are also required to categorize the referent into one of multiple person referent categories. The tool will automatically add each object you create to a list of all the referents in the conversation called the referent list.

✓ Identify each person-referring expression. (Section A.4)

> Your next action is to identify all the person-referring expressions in the utterance. To do this, you will mark a single word in the annotation tool for each expression. For expressions with multiple words, you will need to identify its lexical head using a syntactic analysis of the expression. Most, but not all, of the (lexical heads of) person-referring expressions will already have been automatically identified in the transcript. As this process is not perfect, some additional words may be identified which are not (lexical heads of) person-referring expressions. The entire set of automatically and manually identified words are called markables.

✓ Categorize each markable according to its functional category. (Section A.5)

> Your next action is to categorize each markable into one of multiple functional categories. Here, you will distinguish between different types of referring actions associated with each markable, as well as occasions when automatically-marked markables are not person-referring at all.

✓ Link the referring expressions to their corresponding person referent. (Section A.6)

> Finally, the last action is to link each person-referring expression to its corresponding referent in the referent list. Depending on the type of referent, you may also need to assign some special referring event attributes.

By the end of the conversation, you will have identified all of the people which have been referred to in the conversation and the locations in the conversation where each one was referred to. Each of these steps is described in detail in the instructions below. Throughout the instructions, we will highlight:

⟺ Rules for making annotation decisions

! Exceptions and possible difficulties with applying the scheme

- PERSON REFERENT CATEGORIES
- PRE-DEFINED PERSON REFERENTS
- FUNCTIONAL CATEGORIES
- REFERRING EVENT ATTRIBUTES
- ANNOTATION SHORTCUTS

Your overall approach to the annotation should focus on the speaker's intended meaning and not on explicit literal interpretations. There are many cases where the usual meanings of words do not correspond with how they are used. This means you will need to study the conversational context, looking behind and ahead in the conversation, in order to appreciate the intentions of the speaker. You should rely on introspection, imagery, and common sense. At no time should you try to use technical knowledge of language which is not explicitly part of the annotation procedure.

A.2 Conceptual Analysis of Person-Referring Events

A **person-referring event** is the utterance of a word or phrase (called a person-referring expression) that performs the function of referring to a person (called the person referent). This section describes the concepts you will need to know in order to identify person-referring events. This is not a definition of an explicit sequential procedure, nor is it a description of an explicit action to be performed in the annotation tool. Rather, this is a description of the conceptual analytical goals you should reach before taking any actions in the annotation tool.

A.2.1 Scene analysis of speaker's intended meaning

Your first goal is to determine the meaning of the utterance as intended by the speaker. To do this, rely principally on your intuition to think about the speaker's reasons for speaking, taking into account the conversational context. Use imagery as your principal tool, and perform an analysis of the meaning and speaker's intentions in conceptual terms. It is important not to be distracted by the usual or literal meanings of specific words. Avoid employing any technical knowledge of linguistics at this point.

One helpful analytical tool is the use of imagery to imagine a scene which the speaker's meaning sets up. Allow yourself to create a picture in your mind. It may be a hypothetical or fictional scene which is being imagined or portrayed by the participants. It may be a recurring or typical scene, in which general facts about the world are described. It may be the real physical setting of the conversation. Ultimately, the kinds of scenes set up by speakers is extremely varied and nuanced. There may even be cases where multiple scenes are connected together through the meanings being expressed.

A.2.2 Identifying person referents

Scenes can, of course, involve people. Use your imagery of the scene to determine if there are people involved, either as actors in the scene, viewers of the scene itself, or by virtue of some other relation to it. We call such entities person referents. There may be individuals, collections, or kinds of people in the real world which are brought to mind.

One thing that is important to keep in mind is that there are both implicit and explicit ways that a person can be involved in a scene. Depending on whether the scene is hypothetical, generic, or real, there may be people that are real, imaginary, abstract, or vague observers of the scene. Like the different kinds of scenes, the different kinds of people involved can be extremely varied and nuanced. Almost always, the utterance suggests that the participants themselves are in the scene either explicitly or implicitly.

The following list describes some categories of entities which should be considered person referents.

- A distinct person in the real world
- An imaginary, hypothetical, mythical, or fictitious person
- A human agent, perceiver, or participant in a described event, scene, or fact
- A class, type, or kind of person, or representative thereof
- A specification or description of a person or set or people
- A (possibly vaguely defined) group or collection of any of the above
- The human race as a whole, or a representative thereof

A.2.3 Identifying specific, unambiguous referring

When annotating you will be required to employ the notion of specific, unambiguous referring—a referring event which refers unambiguously to a referent with a specific identity. Whether a referent has a specific identity is part of the interpretation process and is subject to contextual factors. Annotators must therefore use their common sense and understanding of the conversational context to make this distinction. Our requirements for having a specific, unambiguous identity are meant to be very strict — the referring expression must not have any plausible alternate interpretation other than referring to a specific individual or set of individuals.

This can sometimes be tested by inserting necessarily non-specific referring expressions (i.e., one, someone) in place of existing expressions. If the replacement produces a meaning which does not contradict the speaker's intended meaning, then the expression is not specific and unambiguous. Another test involves hypothetically questioning the speaker for clarification. If you had the means to request clarification, would the speaker have a response that was precise (or in the case of a set of people, precisely enumerable)? If not, then the expression is not specific and unambiguous.

A.2.4 Identifying new referents

Your next goal should be to identify whether there are any new person referents, i.e., if any of them have not been previously mentioned in the conversation. Scenes (and elements of those scenes) are often shared between different utterances. That is, the identity of entities can remain unchanged across utterances. This often occurs when speakers require several utterances to get their point across. Or it can occur when a series of connected events is being described. Or it can occur when a scene is brought up for discussion multiple times in a conversation. It can also simply mean that the same real person is discussed on multiple occasions.

You should study the conversational context in order to appreciate when this referent identities are shared between utterances When identifying the person referents in an utterance, the same referent object should be used for those whose identities persist across utterances in this way.

A.2.5 Identifying person-referring expressions

The goals above were about using your intuition to determine the intended meaning of an utterance. The ultimate purpose of the annotation, however, is to match these meanings to observable words. For this reason, you will need to perform a basic syntactic analysis of each utterance. The idea is to match identified person referents with specific words and phrases. We call these phrases person-referring expressions. To determine if a word or phrase is a person-referring expression, the answer to all of the following questions must be "yes".

- ⟺ Is the expression a noun phrase (NP), a possessive determiner, or a possessive pronoun?

  We consider all phrases whose syntactic head is a noun (e.g., a man, the man, most people in Edinburgh), and all phrases which occur in the same syntactic positions as nouns, to be NPs (including possessive NPs, e.g., John's).

- ⟺ Is the word or phrase's primary contribution to utterance meaning the identification, labeling, description, specification, addressing, or referencing of a person referent?

  This scheme considers each of these types of communicative events to be a type of referring.

A.2.6 Common ground

Referring is a collaborative activity. For a referring expression to achieve the speaker's goals for it, addressees must also understand the intended meaning. For the intended meaning of an expression to successfully refer, all relevant parties must attain mutual understanding of the referent. That is, they must have in their minds a notion of the expression's meaning that is similar enough to make the communication work. This is not always straightforward since participants can have different knowledge (of people's names for example) and different perceptions (e.g., whether they can hear or see the speaker). These problems are also a possibility for you, the annotator. Because some of the meetings you will annotate have no video recordings, you will be at a disadvantage in attaining the same level of understanding as the participants. You also do not know the participants names when starting an annotation. Keep these possibilities in mind when annotating.

Some of the categories below relate to this lack of common ground, either between the participants themselves, or between you and the participants.

A.3 Creating Referent Objects

At this point, you are now ready to use your conceptual analysis to perform specific actions in the annotation tool. The first action you will perform is to create a referent object representing each real person (or group of people) who is referred to specifically and unambiguously in the utterance and who was not previously referred to in the conversation. The new object will automatically be given a unique identifier and added to the conversation's referent list.

To create a referent object, use the create referent command. Creating a referent has two steps. First, you assign a nickname or description to the object. If the referent is a single individual and you know the first name of the person, you should use that as the nickname. Otherwise, how you nickname the object is up to you. You can reassign a nickname at any time. Second, you classify the referent into one of the following person referent categories. If a referent does not fall into either of these categories, then it is not real, specific and unambiguous and you do not need to create a referent object representing it.

A.3.1 Person referent categories

- PERSON-REAL-SINGLE

A person referent is PERSON-REAL-SINGLE if the answer to all the following questions is "yes":

⟺ Is it a single individual only?

⟺ Does the individual have a specific, unambiguous identity in the real world?

A name or description is not necessary for having a specific, unambiguous identity.

⟺ Are both of the above known or assumed by the speaker and addressee(s)?

⟺ Are your answers to the questions above the only reasonable interpretation?

For this to be true, the referent must not be replaceable with a generic referent (i.e., one) nor vague or ambiguous in any way.

- PERSON-REAL-MULTIPLE

A person referent is PERSON-REAL-MULTIPLE if the answer to all the following questions is “yes”:

⟺ Is it definitely more than one individual?

⟺ Do each of the individuals meet the requirements for person-real-single?

⟺ Can you, the annotator, precisely enumerate the set of individuals?

⟺ Are your answers to the questions above the only reasonable interpretation?

For this to be true, there must not be any plausible reason for vagueness or ambiguity in determining the members of the set.

A.3.2 Pre-defined referent objects

There are a set of pre-defined referents that are automatically added to the referent list at the beginning of annotation. You can consider these as having been previously mentioned. These are:

- P.1, P.2, etc.
    - a PERSON-REAL-SINGLE referent for each participant
- ALL.PARTICIPANTS
    - a PERSON-REAL-MULTIPLE referent for the set of all participants

! If at any point, you learn the name of a participant, you should nickname the pre-defined referent object using the person’s name.

A.3.3 Annotator-unknown referents

In some cases, when a person referent should be categorized as PERSON-REAL-SINGLE or PERSON-REAL-MULTIPLE, the participants know the referent's identity but you (the annotator) do not. This can be due to two main reasons.

In the first case, your lack of knowledge is a result of not having video. Consider the following example. Someone says he to refer to someone. The participants can clearly identify the referent. However, you can't identify the referent and infer that this is because you can't see the speaker. In this case, you should employ the following special person referent in the annotation tool:

- UNK.NOVIDEO
    - a PERSON-REAL-SINGLE or PERSON-REAL-MULTIPLE referent whose identity is specific but unknown to the annotator due to lack of video

In the second case, a person is referred to by name. The participants know the identity of the referent, but you don't know if the name refers to a participant. Or if you do, you don't know which one. In this case, you should create a new referent for the individual (assuming you can determine that it is a real person). If later during annotation you learn who the name refers to, the annotation tool provides a way for you to merge two referents into a single one.

A.3.4 Exclusion sets

It is often more convenient to specify a PERSON-REAL-MULTIPLE referent by indicating which people to exclude from the set rather than which to include. This is not its own person referent category and such sets must meet the specifications for the PERSON-REAL-MULTIPLE category. You can create such a referent by using the following shortcut and then specifying the participants to exclude. Note however, that this only applies when the set from which you are excluding referents is the set of all participants.

- PERSON-REAL-MULTIPLE (EXCLUSION SET)

A.3.5 Quantified referents

Quantifiers (e.g., many, no, few, each, some, every, lots, two) are usually used to specify a subset of some larger set. For example, in the phrase a few of you, the word you on its own specifies a set while adding a few of specifies a subset of that set. In some cases (e.g., all, every, each) the specified subset is equivalent to the larger set. In others (e.g., no, none of) the subset is empty. These cases cause difficulty because they can be considered to be two embedded referential expressions.

As with all person-referring expressions, it is important when annotating quantified expressions that you always first consider the meaning of the entire phrase. If the meaning of the entire phrase is PERSON-REAL-SINGLE or PERSON-REAL-MULTIPLE, then it should be annotated without regard for whether it is quantified or not. For example, all of us referring to all the participants should be annotated as PERSON-REAL-MULTIPLE referring to ALL.PARTICIPANTS. However, the following exception applies to some quantified expressions.

! If the entire quantified phrase does not refer to a PERSON-REAL-SINGLE or PERSON-REAL-MULTIPLE referent, you should annotate the expression according to the larger set from which the quantifier is specifying a subset and label the referring event with the following attribute. (The attribute should be labeled as part of the final step in annotating the referring event.)

- ATTR-QUANTIFIER-SUPERSET

The meaning of the larger set is usually dependent on the noun which the quantifier is modifying. For example, a few of us, might refer to an unspecified subset of the larger set of all participants, where us on its own refers to all participants. In this case, you should annotate the expression as referring to the PERSON-REAL-MULTIPLE ALL.PARTICIPANTS but also assign the ATTR-QUANTIFIER-SUPERSET attribute to the referring event (assigning referring event attributes happens as the last step in the annotation of each markable).

Note also that negatively quantified person-referring expressions (expressions referring to the empty set) should also be handled in this way. However, they should not be confused with negated predicates, e.g., You are not a lawyer. Such examples are not considered to be quantified referents—the not should be considered as modifying the VP instead of the NP.

A.4 Identifying Person-Referring Expressions

A.4.1 Markables and person-referring expressions

A person-referring expression is an occurrence of a word or phrase which performs the function of referring to a person referent. Every occurrence of a person-referring expression is a markable, i.e., an expression which must be annotated. However, not all markables are person-referring expressions. There is a subset of markables, called the closed-class markables, which are identified simply according to whether they are an occurrence of a pre-defined list of word forms. Not every occurrence of these word forms is a person-referring expression. However, most person-referring expressions are closed-class markables (and vice-versa), so the two sets are largely intersecting. The set of all markables is defined as the union of these two sets. The closed-class set of word forms are:

- anybody, anyone, everybody, everyone, he, her, hers, herself, him, himself, his, I, me, mine, my, myself, nobody, our, ours, ourselves, she, somebody, someone, their, theirs, them, themselves, they, us, we, you, your, yours, yourself, yourselves

A.4.2 Identifying lexical heads

The second action you will perform, after adding any new referents, is to identify all the open-class person-referring expressions in the utterance. Closed-class markables will already be identified and displayed in bold in the annotation tool. Your job is to use the toggle markable command to add the open-class referring expressions to the set of bold words. However, some of the open-class expressions contain multiple words. Multi-word referring expressions should always be analyzed according to the meaning of the entire phrase (except some quantified expressions, as described above). However, you are not required to annotate the actual extent of the phrase. Instead, to identify a multi-word referring expression in the annotation tool, you simply need to identify its lexical head. Henceforth, lexical heads will be indicated by underlining in examples, and the extent of any multi-word referring expressions will be indicated by brackets.

To identify the lexical head of a multi-word referring expression, the general rule is to identify its last head noun or the pronoun (with some exceptions, see below). Most noun phrases are a sequence of determiners, pre-modifiers, nouns/pronouns, and post-modifiers, with the noun/pronoun being the only non-optional component (though in rare cases it may be absent). There are four main cases to consider:

- In cases where the phrase contains a single noun/pronoun, it is the lexical head.
- In cases when the phrase contains multiple nouns or a proper name with multiple words, you should choose the last noun or last word in the proper name as the lexical head.
- In the rare case where there is no noun or pronoun, there will most likely be an adjective, quantifier, or other pre-modifier acting as the head of the phrase (the missing noun would occur just after the acting head), e.g., the smartest, a few, you two in the back. In these cases, label the adjective, quantifier, or other pre-modifier as the lexical head.
- In reciprocal phrases like each other and one another, select the final word in the phrase as the lexical head.

A.5 Classifying Markables into Functional Categories

After identifying all the markables in the utterance, your next task is to annotate each markable. To do this you will use the label markable command. Issuing this command will take you through a series of annotation choices. The first choice you will make is to classify the markable according to its functional category. Functional categories relate to the communicative purpose of a word or phrase. To label the functional category, proceed sequentially through the following list of categories. If the markable meets the specification for a category, assign it and skip the remaining categories. The functional categories are as follows:

- FUNC-FILLER
- FUNC-NON-PERSONREF
- FUNC-PERSONREF-TROUBLE
- FUNC-PERSONREF-VOCATIVE
- FUNC-PERSONREF-INTRODUCTION
- FUNC-PERSONREF-DEFAULT

Note that if that procedure for identifying person-referring expressions is used correctly, it is impossible for any markables that are not closed-class markables to be in the following categories: FUNC-FILLER, and FUNC-NON-PERSONREF. These categories are therefore only offered by the tool for closed-class markables. The remaining categories are all considered to be types of person-referring.

- FUNC-FILLER (closed-class markables only)

To qualify for this category, all of the following questions must be answered "yes".

⟺ Does the markable occur in one of the following phrases?

> you know, you see, I mean, let's see, let's say

⟺ Is the phrase's primary purpose not to motivate a response from the addressee?

⟺ Does the phrase occur as a parenthetical?

> Parentheticals are syntactically isolated phrases, meaning that removing them would not disrupt the syntax of the utterance.

⟺ Does the phrase not contribute any propositional meaning to the utterance?

> Uses which suggest any kind of explicit, literal interpretation should be considered as contributing propositional meaning to the utterance.

- FUNC-NON-PERSONREF (closed-class markables only)

⟺ Is the markable not a person-referring expression (or the lexical head of one)?

- FUNC-PERSONREF-TROUBLE

To qualify for this category, all of the following questions must be answered "yes".

⟺ Is the markable unsuccessful in establishing shared understanding, between the speaker and addressee(s), of the identity of the person referent?

⟺ Is that lack of success evidenced by multiple attempts to make the reference, either by repetition, rephrasing, or clarification?

- ⟺ Does the trouble establishing common ground concern the person referent specifically?
- ⟺ Every instance of referring which is unsuccessful should be classified to this category. But keep in mind there is usually a final reference that establishes shared understanding which should not be classified to this category. Each occurrence of this category will be automatically linked to the person-other referent other.

- FUNC-PERSONREF-VOCATIVE

To qualify for this category, all of the following questions must be answered “yes”.

- ⟺ Is the markable not the word you?
- ⟺ Is the principal purpose of the markable to address a participant, gain their attention, or select them as the next speaker or addressee?

- FUNC-PERSONREF-INTRODUCTION

To qualify for this category, all of the following questions must be answered “yes”.

- ⟺ Does the markable refer to a participant?
- ⟺ Does the markable refer to the participant by name?
- ⟺ Does the markable occur in an utterance where the individual is being introduced (or introducing themselves) to the other participants?
- ⟺ By introduction, we mean the social activity of “meeting” or “being introduced", as is done between individuals who have not met before or for those who do not know the participant’s name.

- FUNC-PERSONREF-DEFAULT

- ⟺ Is the markable a person-referring expression (or the lexical head of one), but not FUNC-PERSONREF-TROUBLE, FUNC-PERSONREF-VOCATIVE, nor FUNC-PERSONREF-INTRODUCTION?

A.6 Labeling Referential Properties

A.6.1 Linking markables to referents

The second step in labeling markables is to link markables that are person-referring to the appropriate person referents. This includes all markables categorized as any of the following: FUNC-PERSONREF-TROUBLE, FUNC-PERSONREF-VOCATIVE, FUNC-PERSONREF-INTRODUCTION, and FUNC-PERSONREF-DEFAULT. Closed-class markables which are not person-referring do not need to be labeled.

The annotation tool provides some shortcuts for selecting referents more easily. These are not referents themselves, but rather assist you in selecting referents based on a specification. The shortcuts are as follows:

- SPEAKER
    - Selects the speaker of the currently selected word as the referent.

A.6.2 Membership attributes

The final step in labeling markables is to label their membership attributes. Not all person-referring expressions refer specifically and unambiguously to real, single or multiple person referents. Usually, the conversational context and the type of scene which the speaker sets up suggests or implies that real individuals are either included or excluded from the set of individuals being referring to. The principal dimension along which individuals are included or excluded from such underspecified referring is the one that distinguishes the speaker, the addressee(s), and others. This dimension parallels the distinction between the basic meanings of first, second, and third person pronouns. However, most pronouns, especially "we" and "you", can have extremely varied meanings which are sometimes contrary to their basic meanings.

You goal is to label the membership attributes of any referring events that do not refer unambiguously to a specific referent. To do this, you must look past the usual meanings of words in order to assess whether the speaker is implying that certain individuals are being referred to (or not). For example, if a person is describing their own experiences or habits, but is using the word you with the effect of de-personalizing or genericizing the events, there will be an implied aspect of self-inclusion in the referent in addition to the usual addressee-inclusion associated with the word you. Membership attributes can also be used to distinguish uses of we which include the addressee from those which exclude the addressee (in addition to the speaker inclusion normally implied by first-person pronouns).

The following is the list of membership attributes. Assigning an attribute implies that the associated referent is included in the reference. Leaving the attribute unassigned means that exclusion is implied or suggested.

- ATTR-SPEAKER
    - Assigned if the membership of a referent suggests or implies inclusion of the speaker of the utterance (unassigned if exclusion is implied).
- ATTR-ADDRESSEE
    - Assigned if the membership of a referent suggests or implies inclusion of the principal addressee(s) of the utterance (unassigned if exclusion is implied).
- ATTR-OTHER
    - Assigned if the membership of a referent suggests or implies inclusion of individuals other than the speaker and addressee(s) (unassigned if exclusion is implied).

A.7 Further Discussion

This section presents further details on difficult or nuanced aspects of the annotation scheme and method.

A.7.1 Unknown participant names

Usually you do not know the participant's names when starting to annotate a conversation. Rather, you will learn their names by listening to the conversation. This means that you may not know the identity of a named participant (i.e., the person referent corresponding to a mentioned name). When this happens, you should create a new person referent and assign them an appropriate nickname. Later on, when you learn the identity of the named participant, you can use the merge referent action to cause all references to either of two referents to be references to a single referent.

Note that you might be able to avoid using the merge referent feature by searching the dialogue for the name. This could allow you to learn the identity of the named participant in advance of assigning any referents.

Another complication is when references are made to the complement of a set of participants using names with unknown identities (e.g., [everyone except John], where the identity of John is unknown). The solution to this is to create a new person referent and assign them an appropriate nickname (i.e., "John"). Then create a negatively defined set which excludes John. When you finally learn the identity of John, you can then merge the John referent with his known identity (as above), and the negatively defined set will then be correct.

A.7.2 Types of referring

This annotation scheme considers a broad range of person-referring events. A person-referring event is the occurrence of an NP which identifies, labels, describes, specifies, addresses, or references a person referent as its primary contribution to the speaker's intended meaning. Any one of these specific types of communicative events should be considered person-referring. Some of these types correspond to the functional categories.

Addressing and Vocatives

An addressee is the person or set of people to which a speaker directs their utterance. The most common way to refer to an addressee is with the word you. One other type of addressing, which we distinguish with the FUNC-PERSONREF-VOCATIVE category, directly addresses a person using their name, e.g., "Hey, [John]. Can [you] hand [me] the book?" In this case, the addressing use John is FUNC-PERSONREF-VOCATIVE while you and me are FUNC-PERSONREF-DEFAULT. A person's name is usually used in this way to gain their attention, single them out as the intended addressee, or to select them as the next speaker.

Describing, Labeling, Specifying, and Naming

Utterances which use copular verbs like be and seem present a unique problem to defining reference. Such utterances often both identify as well as either describe, label, specify or name a person, e.g., "[I] am [John], [the project manager]." In these kinds of utterances, any NPs which performs any of these functions should be labeled. In this particular case, the John is FUNC-PERSONREF-INTRODUCTION because this utterance introduces a person to the others and is a mention of that person's name. An expression which describes the attributes of an individual (e.g., a lawyer in the utterance I am a lawyer) should be annotated as referring to that individual, even if the description is negated (e.g., I am not a lawyer).

A.7.3 References to names and words

The use of people's names and words which normally refer to people can sometimes not be person-referring but instead refer to the name or word itself. Remember to rely on imagery to determine if the referent is actually a person. References to words or names usually do not evoke an image of a person. One way to think about this is to determine if it would be reasonable to place quotation marks around the word in a transcription.

A.7.4 Syntax, multi-word expressions, and nesting

A general rule is that referring expressions should be annotated according to the meaning conveyed by all of the words in the phrase. This means that one should consider any determiners (articles, possessives, demonstratives, numerals or quantifiers), pre-modifiers (adjectival or nominal), or post-modifiers (relative clauses, non-finite clauses, or prepositional phrases) and their contribution to the meaning of the phrase. However, a person-referring expression can sometimes syntactically dominate another. As long as each expression does not have the same lexical head, they should be annotated independently as separate referring expressions. If the two expressions have the same lexical head (this should only be the case with quantified expressions), then follow the rules for annotating quantified expressions. Exceptions to this rule include lists and coordinated NPs (expressions containing coordinators like and and or, e.g., John and Jim). These should not be considered as single NPs and should therefore not be marked as a single expression. Instead, their components should be marked as separate person-referring expressions.

Another general rule is that all closed-class markables must be annotated, but when they occur as the lexical head of a super-ordinate open-class markable, they should be annotated according to the meaning of the superordinate phrase (an open-class markable is a markable which is not closed-class). There are two major cases when a closed-class markable occurs within a super-ordinate open-class markable. In the first, the closed-class markable coincides with the lexical head of the super-ordinate (e.g., [one of you]). These cases should be handled according to the rules for quantified expressions. The other case is where the lexical head of the superordinate phrase and the closed-class markable are not the same (e.g., [you guys], [my brother]). In this case, both words need to be annotated. The you should be marked as FUNC-NON-PERSONREF since it is being used as a determiner and is not a noun phrase. The guys should then be marked as the lexical head of you guys, and should be annotated according to the meaning of the entire phrase you guys. The my should be annotated according to the person referent holding the possession relation, and the brother should be annotated according to the meaning of the entire phrase my brother.

A.7.5 Possessives

Possessives express an ownership relation between a person and something possessed. For non-closed-class markables, this is usually marked with the genitive suffix 's. In these cases, the suffix should already be tokenized separately, and the word or phrase referring to the person holding the possession relationship (which should be considered an NP in this case) should be annotated. The set of closed-class markables also includes possessive pronouns (e.g., mine, yours) and determiners (e.g., my, your). These should be annotated as if they were NPs referring to the person holding the possession relationship.

A.7.6 Metonymy

Metonymy is a figure of speech where an expression refers to something related to its usual referent (e.g., "I 'm on channel three"). In this case the person speaking is not literally "on" channel three, but rather the audio signal carrying their voice is on channel three. These cases should be considered as referring to the person.

A.7.7 Organizations vs. groups of people

It is common to confuse groups of people with organizations. Organizations often behave like people by making decisions or taking actions. People also often act on behalf of organizations. The rule of thumb for dealing with this problem is to use imagery and to decide whether the primary image that comes to mind, when considering its context of use in the utterance, is a person, group of people, or kind of person. If it is not and is instead primarily an image of an institution or organizational body, then it is not a person referent.

A.7.8 Wh- pronouns

Do not annotate wh- pronouns as markables.

A.7.9 Transcription problems

Some utterances are either incorrectly transcribed, or the audio is censored by a beep. Please just ignore these and do not annotate them.

A.8 Using the Annotation Tool

The annotation tool requires you to use only the keyboard. You will not need to use the mouse. Each word is displayed in chronological sequence. Words are assigned to different rows depending on who spoke them, with each row's speaker labelled at the margins. The current playback position of the media file is displayed above the word rows. The current position of the cursor is displayed below them. Markables are indicated in bold. Markables which have been assigned to a functional category are indicated with an underline.

Moving the cursor	
PREV/NEXT MARKABLE	Move the cursor to the previous/next markable, whether it is annotated or not.
PREV/NEXT UNLABELED MARKABLE	Move the cursor to the previous/next unlabeled markable.
PREV/NEXT WORD	Move the cursor to the previous/next word.
FOCUS ON PLAYHEAD	Move the cursor so it is aligned with the current location of media playback.
JUMP TO TIME	Move the cursor to a particular time (in seconds) in the conversation.
FORWARD/BACKWARD SEARCH	Move the cursor to the next/previous occurrence of a given string. The search checks the word itself first (exact matches only), then the functional category label (partial matches), and finally the referent nickname (partial matches).

Media actions	
PLAY/PAUSE	Stop playback if it is playing, or start playback if it is stopped.
PLAY AT CURSOR	Begin playback at the specified number of seconds (0 through 9) prior to the cursor.
RELOAD MEDIA	Occasionally, the media actions will become unresponsive. Use this key to reload the media file into the tool.

Annotating markables	
TOGGLE MARKABLE	This command identifies which words are the lexical heads of markables. The closed-class words cannot be toggled.
LABEL MARKABLE (FUNC-PERSONREF-DEFAULT)	This command assigns the markable to the functional category func-personref-default, and then prompts you to assign a referent.
LABEL MARKABLE (OTHERS)	This command prompts you to choose which functional category (other than func-personref-default) to assign to a markable. If you choose one of the referring categories, you are then prompted to assign a referent.

Annotating referents	
CREATE REFERENT	This command takes you through the process of adding a new referent to the referent list.
NICKNAME REFERENT	This command allows you to re-assign the nickname of an existing referent in the referent list.
EDIT REFERENT	This command allows you to re-specify the set of individuals in a person-real-multiple referent in the referent list.
MERGE REFERENTS	This command re-assigns every markable that refers to a chosen referent so that it refers to another chosen referent. The former is said to be destroyed and the other merged to. This is a useful command when references to named individuals occur prior to the annotator knowing which individual (identity) that name refers to. You cannot destroy a participant referent. Therefore, you should choose to destroy any new temporary referents you have created and merge to the automatically created participant referent.

Display actions	
SCROLL	This moves the display from right to left, without changing the cursor location.
TOGGLE WIDE VIEW	This provides a zoomed-out view of the conversation.
CLEAR SCREEN	The window will sometimes get filled with odd characters. Use this command to refresh the display.

Other actions	
CANCEL	Cancels any action which is cancelable.
QUIT	Quit the program (saves a backup file).
SAVE	Save the annotation file (also saves a backup).

Appendix B

Simplified Annotation Instructions

Perform the following for each utterance in the conversation.

✓ Add any new specific person referents to the referent list.

PERSON-REAL-SINGLE

- *single, specific, unambiguous identity in the world*

PERSON-REAL-MULTIPLE

- *multiple, precisely enumerable, specific identities in the world*

✓ Mark each person-referring expressions by identifying its lexical head.

✓ Assign every markable to a functional category.

FUNC-FILLER *closed-class only*

- *you know, you see, I mean, I guess, let's see, let's say*
- *does not motivate response; parenthetical; no propositional meaning*

FUNC-NON-PERSONREF *closed-class only*

- *not person-referring*

FUNC-PERSONREF-TROUBLE

- *part of occurrence of trouble establishing common ground between participants*

FUNC-PERSONREF-VOCATIVE

- *not you; used to address, gain attention, or select next speaker*

FUNC-PERSONREF-INTRODUCTION

- *a participant's name; used in utterance that introduces a participant*

otherwise FUNC-PERSONREF-DEFAULT

✓ Link each person-referring expression to its appropriate referent in the referent list.

✓ Label any non-specific referring events for membership attributes

ATTR-SPEAKER-INCL/EXCL

- *does the reference imply inclusion/exclusion of the speaker*

ATTR-ADDRESSEE-INCL/EXCL

- *does the reference imply inclusion/exclusion of the addressee(s)*

ATTR-OTHER-INCL/EXCL

- *does the reference imply inclusion/exclusion of anyone but the speaker and addressee(s)*

Appendix C

List of Stop Words

- n’t
- ’m
- ’s
- ’re
- ’ll
- a
- about
- above
- across
- after
- afterwards
- again
- against
- all
- almost
- alone
- along
- already
- also
- although
- always
- am
- among
- amongst
- amount
- an
- and
- another
- any
- anyhow
- anyone
- anything
- anyway
- anywhere
- are
- around
- as
- at
- back
- be
- became
- because
- become
- becomes
- becoming
- been
- before
- beforehand
- behind
- being
- below
- beside
- besides
- between
- beyond
- both
- bottom
- but
- by
- call
- can
- cannot
- cant
- co
- con
- could
- couldnt
- cry
- de
- do
- done
- down
- due
- during
- each
- eg
- eight
- either
- eleven
- else
- elsewhere
- empty
- enough
- etc
- even
- ever
- every
- everyone
- everything
- everywhere
- except
- few

- fifteen
- first
- five
- for
- former
- formerly
- forty
- found
- four
- from
- front
- full
- further
- get
- give
- go
- had
- has
- hasnt
- have
- he
- hence
- her
- here
- hereafter
- hereby
- herein
- hereupon
- hers
- herself
- him
- himself
- his
- how
- however
- hundred
- i
- ie
- if
- in
- inc
- indeed
- interest
- into
- is
- it
- its
- itself
- keep
- last
- latter
- latterly
- least
- less
- ltd
- made
- many
- may
- me
- meanwhile
- might
- mill
- mine
- more
- moreover
- most
- mostly
- move
- much
- must
- my
- myself
- name
- namely
- neither
- never
- nevertheless
- next
- nine
- no
- nobody
- none
- noone
- nor
- not
- nothing
- now
- nowhere
- of
- off
- often
- on
- once
- one
- only
- onto
- or
- other
- others
- otherwise
- our
- ours
- ourselves
- out
- over
- own
- part
- per
- perhaps
- please
- put
- rather
- re
- same
- see
- seem
- seemed
- seeming
- seems
- several
- she
- should
- show
- side
- since
- six
- sixty
- so
- some
- somehow
- someone
- something
- sometime
- sometimes
- somewhere
- still
- such
- take
- ten
- than

- that
- the
- their
- them
- themselves
- then
- thence
- there
- thereafter
- thereby
- therefore
- therein
- thereupon
- these
- they
- thick
- thin
- third
- this
- those
- though
- three
- through
- throughout
- thru
- thus
- to
- together
- too
- top
- toward
- towards
- twelve
- twenty
- two
- un
- under
- until
- up
- upon
- us
- versa
- very
- via
- vice
- was
- we
- well
- were
- what
- whatever
- when
- whence
- whenever
- where
- whereafter
- whereas
- whereby
- wherein
- whereupon
- wherever
- whether
- which
- while
- whither
- who
- whoever
- whole
- whom
- whose
- why
- will
- with
- within
- without
- would
- yes
- yet
- you
- your
- yours
- yourself
- yourselves

Bibliography

Jan Alexandersson. *Hybrid Discourse Modeling and Summarization for a Speech-to-Speech Translation System*. PhD thesis, Universität des Saarlandes, 2003.

Jan Alexandersson, Bianka Buschbeck-Wolf, Tsutomu Fujinami, Michael Kipp, Stephan Koch, Elisabeth Maier, Norbert Reithinger, Birte Schmitz, and Melanie Siegel. Dialogue acts in VERBMOBIL-2. Technical Report 226, DFKI Saarbrucken, 1998.

Jan Alexandersson, Peter Poller, Michael Kipp, and Ralf Engel. Multilingual summary generation in a speech-to-speech translation system for multilingual dialogues. In *Proceedings of the First International Conference on Natural Language Generation*, pages 148–155. Association for Computational Linguistics, 2000.

James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, and Yiming Yang. Topic detection and tracking pilot study: Final report. In *Proceedings of the 1998 DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

James Allen and Raymond Perrault. Analyzing intention in utterances. *Artificial Intelligence*, 15(3):143–178, 1980.

Jens Allwood. An activity based approach to pragmatics. In Harry Bunt and B. Black, editors, *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*. John Benjamins, 2000.

AMI Consortium. *Guidelines for Dialogue Act and Addressee Annnotation*, 2008a. Downloaded from corpus.amiproject.org on Nov 17, 2008.

AMI Consortium. *Abstractive Hand Summaries Guidelines (Scenario Data)*, 2008b. Downloaded from corpus.amiproject.org on Nov 17, 2008.

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

John L. Austin. *How to Do Things with Words*. Clarendon, 1962.

Satanjeev Banerjee and Alexander Rudnicky. Detecting the noteworthiness of utterances in human meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 71–78. Association for Computational Linguistics, 2009.

Satanjeev Banerjee, Carolyn Rosé, and Alex Rudnicky. The necessity of a meeting recording and playback system, and the benefit of topic-level annotations to meeting browsing. In *Proceedings of the 10th International Conference on Human-Computer Interaction (CHI)*, 2005.

Adrian Bangerter, Herbert H. Clark, and Anna R. Katz. Navigating joint projects in telephone conversations. *Discourse Processes*, 37(1):1–23, 2004.

Doug Beeferman, Adam Berger, and John D. Lafferty. Statistical models for text segmentation. *Machine Learning*, 34(1-3):177–210, 1999.

Douglas Biber. *Dimensions of register variation: a cross-linguistic comparison*. Cambridge University Press, 1995.

Douglas Biber. A register perspective on grammar and discourse: Variability in the form and use of english complement clauses. *Discourse Studies*, 1(2):131–150, 1999.

Douglas Biber, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. *Longman Grammar of Spoken and Written English*. Longman, 1999.

Sasha Blair-Goldensohn and Kathleen McKeown. Integrating rhetorical-semantic relation models for query-focused summarization. In *Proceedings of the 6th Document Understanding Conference (DUC2006)*, 2006.

Harold Borko and Charles L. Bernier. *Abstracting Concepts and Methods*. Academic Press, 1975.

Thorsten Brants and Alex Franz. Web 1T 5-gram, Version 1. Linguistic Data Consortium, 2006. LDC2006T13.

Michael E. Bratman. *Intentions, Plans, and Practical Reason*. CSLI Publications, 1987.

Gillian Brown and George Yule. *Discourse Analysis*. Cambridge University Press, 1983.

Trung H. Bui, Matthew Frampton, John Dowding, and Stanley Peters. Extracting decisions from multi-party dialogue using directed graphical models and semantic similarity. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, 2009.

Harry Bunt. Dimensions in dialogue act annotation. In *Proceedings of the Language Resources and Evaluation Confrerence*, 2006.

Harry Bunt. DIT++ annotation scheme, 2008. Release 3, version 2.

Harry Bunt and William Black, editors. *Abduction, belief, and context in dialogue: studies in computational pragmatics*. John Benjamins, 2000.

Harry Bunt, Jan Alexandersson, Jean Carletta, Jae-Woong Choe, Alex Chengyu Fang, Koiti Hasida, Kiyong Lee, Volha Petukhova, Andrei Popescu-Belis, Laurent Romary, Claudia Soria, and David Traum. Towards an ISO standard for dialogue act annotation. In *Proceedings of the Language Resources and Evaluation Confrerence*, 2010.

Susanne Burger, Victoria MacLaren, and Hua Yu. The ISL Meeting Corpus: The impact of meeting type on speech style. In *Proceedings of the 7th International Conference on Spoken Language Processing*, 2002.

Lou Burnard. *Reference Guide for the British National Corpus (XML Edition)*, 2007.

Miriam Butt. The light verb jungle. *Harvard Working Papers in Linguistics*, 9:1–49, 2003.

Sasha Calhoun, Malvina Nissim, Mark Steedman, and Jason Brenier. A framework for annotating information structure in discourse. In *Proceedings of the ACL Workshop: Frontiers in Corpus Annotation II: Pie in the Sky*, 2005.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the Annual ACM SIGIR Conference*, 1998.

Jean Carletta. Unleashing the killer corpus: Experiences in creating the multi-everything AMI meeting corpus. *Language Resources and Evaluation Journal*, 41:181–190, 2007.

Jean Carletta, Amy Isard, Stephen Isard, Jacqueline Kowtko, Gwnyeth Doherty-Sneddon, and Anne Anderson. *HCRC Dialogue Structure Coding Manual*, 1996.

Jean Carletta, Stephen Isard, Anne H. Anderson, Gwyneth Doherty-Sneddon, Amy Isard, and Jacqueline C. Kowtko. The reliability of a dialogue structure coding scheme. *Computational Linguistics*, 23(1):13–31, 1997.

Wallace L. Chafe. Should computers write spoken language? In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 27–28. Association for Computational Linguistics, 1980a.

Wallace L. Chafe, editor. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production*, volume 3 of *Advances in Discourse Processes*. Ablex, 1980b.

Wallace L. Chafe. *Discourse, Consciousness, and Time: The Flow and Displacement of Conscious Experience in Speaking and Writing*. University of Chicago Press, 1994.

Berlin Chen and Yi-Ting Chen. Word topical mixture models for extractive spoken document summarization. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*, pages 52 –55, 2007.

Berlin Chen and Yi-Ting Chen. Extractive spoken document summarization for information retrieval. *Pattern Recognition Letters*, 29(4):426 – 437, 2008.

Yi-Ting Chen, Shih-Hsiang Lin, Hsin-Min Wang, and Berlin Chen. Spoken document summarization using relevant information. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 189 –194, 2007.

Yi-Ting Chen, Berlin Chen, and Hsin-Min Wang. A probabilistic generative framework for extractive broadcast news speech summarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(1):95 –106, 2009.

Freddy Y. Y. Choi. Advances in domain independent linear text segmentation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 26–33, 2000.

Freddy Y. Y. Choi, Peter Wiemer-Hastings, and Johanna Moore. Latent semantic analysis for text segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 109–117, 2001.

Alexander Clark and Andrei Popescu-Belis. Multi-level dialogue act tags. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, 2004.

Herbert H. Clark. *Arenas of Language Use*. University of Chicago Press, 1992.

Herbert H. Clark. *Using Language*. Cambridge University Press, 1996.

Herbert H. Clark and Susan E. Brennan. Grounding in communication. In Lauren B. Resnick, John M. Levine, and Stephanie D. Teasley, editors, *Socially Shared Cognition*, pages 127–149. 1991.

James Clarke and Mirella Lapata. Discourse constraints for document compression. *Computational Linguistics*, 36:411–441, 2010.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46, 1960.

Philip R. Cohen and Raymond Perrault. Elements of a plan-based theory of speech acts. *Cognitive Science*, 3:177–212, 1979.

Jeff Conklin. Dialog mapping: Reflections on an industrial strength case study. In *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*. Springer, 2003.

Mark Core and James Allen. Coding dialogues with the DAMSL annotation scheme. In D. Traum, editor, *Proceedings of the 1997 AAAI Fall Symposium on Communicative Action in Humans and Machines*, pages 28–35, 1997.

Jim Cowie and Wendy Lehnert. Information extraction. *Communications of the ACM*, 39 (1):80–91, January 1996.

Jim Cowie and Yorick Wilks. Information extraction. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *Handbook of Natural Language Processing*, pages 249–269. 2000.

Anita Cremers and Bart Hilhorst. What was discussed by whom, how, when and where? In *Proceedings of the 11th International Conference on Human-Computer Interaction (HCI International)*, 2005.

K. H. Davies, R. Biddulph, and S. Balashek. Automatic speech recognition of spoken digits. *Journal of the Acoustical Society of America*, 24:637–642, 1952.

Marie-Catherine de Marneffe, Bill MacCartney, Trond Grenager, Daniel Cer, Anna Rafferty, and Christopher D. Manning. Learning to distinguish valid textual entailments. In *Proceedings of the Second PASCAL Challenges Workshop*, 2006a.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. Generating typed dependency parses from phrase structure parses. In *Proceedings of the Language Resources and Evaluation Confrerence*, pages 562–569, 2006b.

Larry Denenberg, H. Gish, M. Meeter, T. Miller, J. R. Rohlicek, W. Sadkin, and M. Siu. Gisting conversational speech in real-time. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 131–134, 1993.

Nigel Dewdney, Carol VanEss-Dykema, and Richard MacMillan. The form is the substance: classification of genres in text. In *Proceedings of the Workshop on Human Language Technology and Knowledge Management*, 2001.

Inderjit S. Dhillon, Subramanyam Mallela, and Dharmendra S. Modha. Information-theoretic co-clustering. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 89–98, 2003.

Rajdip Dhillon, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. Meeting recorder project: Dialog act labeling guide. Technical Report TR-04-002, ICSI, February 2004.

Barbara Di Eugenio, Pamela W. Jordan, and Liina Pylkkännen. *The COCONUT project: dialogue annotation manual, DRAFT*, 1997.

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. The Automatic Content Extraction (ACE) program tasks, data, and evaluation. In *Proceedings of the Language Resources and Evaluation Confrerence*, 2004.

Patrick Ehlen, Raquel Fernández, and Matthew Frampton. Designing and evaluating meeting assistants, keeping humans in mind. In *Proceedings of the Workshop on Machine Learning and Multimodal Interaction*, 2008.

Jacob Eisenstein and Regina Barzilay. Bayesian unsupervised topic segmentation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 334–343, 2008.

Brigitte Endres-Niggemeyer, Elisabeth Maier, and Alexander Sigel. How to implement a naturalistic model of abstracting: Four core working steps of an expert abstractor. *Information Processing & Management*, 31(5):631 – 674, 1995.

Raquel Fernández, Matthew Frampton, John Dowding, Anish Adukuzhiyil, Patrick Ehlen, and Stanley Peters. Identifying relevant phrases to summarize decisions in spoken meetings. In *Proceedings of the Interspeech Conference*, 2008a.

Raquel Fernández, Matthew Frampton, Patrick Ehlen, Matthew Purver, and Stanley Peters. Modelling and detecting decisions in multi-party dialogue. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, 2008b.

Charles J. Fillmore. *Lectures on Deixis*. CSLI Publications, 1997.

Michael Finke, Maria Lapata, Alon Lavie, Lori Levin, Laura Mayfield Tomokiyo, Thomas Polzin, Klaus Ries, Alex Waibel, and Klaus Zechner. CLARITY: inferring discourse structure from speech. In *Proceedings of the AAAI Spring Symposium on Applying Machine Learning to Discourse Processing*, 1998.

Jon Fiscus, George Doddington, John Garofolo, and Alvin Martin. NIST's 1998 topic detection and tracking evaluation (TDT2). 1998. Downloaded from www.itl.nist.gov on Sep. 22, 2010.

Matthew Frampton, Raquel Fernández, Patrick Ehlen, Anish Adukuzhiyil, and Stanley Peters. Leveraging minimal user input to improve targeted extraction of action items. In *Proceedings of Semdial*, 2008.

Matthew Frampton, Raquel Fernández, Patrick Ehlen, Mario Christoudias, Trevor Darrell, and Stanley Peters. Who is "you"? combining linguistic and gaze features to resolve second-person references in dialogue. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 2009a.

Matthew Frampton, Jia Huang, Trung Huu Bui, and Stanley Peters. Real-time decision detection in multi-party dialogue. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2009b.

Yasuhisa Fujii, Norihide Kitaoka, and Seiichi Nakagawa. Automatic extraction of cue phrases for important sentences in lecture speech and automatic lecture speech summarization. In *Proceedings of the Interspeech Conference*, 2007.

Sadaoki Furui, Yousuke Shinnaka, Tomonori Kikuchi, and Chiori Hori. Speech-to-text and speech-to-speech summarization of spontaneous speech. *IEEE Transactions on Speech and Audio Processing*, 12(4):401 – 408, 2004.

Michel Galley. A skip-chain conditional random field for ranking meeting utterances by importance. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2006.

Michel Galley, Kathleen McKeown, Eric Fosler-Lussier, and Hongyan Jing. Discourse segmentation of multi-party conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 562–569, 2003.

Michel Galley, Kathleen McKeown, Julia Hirschberg, and Elizabeth Shriberg. Identifying agreement and disagreement in conversational speech: Use of Bayesian networks to model pragmatic dependencies. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2004.

Nikhil Garg, Benoit Favre, Korbinian Reidhammer, and Dilek Hakkani-Tür. ClusterRank: a graph based method for meeting summarization. In *Proceedings of the Interspeech Conference*, 2009.

John S. Garofolo, Ellen M. Voorhees, Vincent M. Stanford, and Karen Spärck Jones. Trec-6 1997 spoken document retrieval track overview and results. In *Proceedings of the Text Retrieval Conference*, 1997.

Maria Georgescul, Alexander Clark, and Susan Armstrong. An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 144–151, 2006.

Sebastian Germesin and Theresa Wilson. Agreement detection in multiparty conversation. In *Proceedings of the International Conference on Multimodal Interfaces*, 2009.

Dan Gillick, Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tur. A global optimization framework for meeting summarization. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2009.

John J. Godfrey, Edward Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 517–520, 1992.

Erving Goffman. *Frame analysis: An essay on the organization of experience*. Harper & Row, 1974.

Erving Goffman. Footing. *Semiotica*, 25:1–29, 1979. Reprinted in E. Goffman. Forms of Talk. 1981. pp. 124–159. Oxford: Blackwell.

Charles Goodwin. Participation, stance and affect in the organization of activities. *Discourse & Society*, 18(1):53–73, 2007.

H. Paul Grice. Meaning. *The Philosophical Review*, 66:377–388, 1957.

H. Paul Grice. Utterer's meaning and intentions. *The Philosophical Review*, 78:147–177, 1969.

H. Paul Grice. Logic and conversation. In D. Davidson and G. Harman, editors, *The Logic of Grammar*, pages 64–75. Dickenson, 1975.

Ralph Grishman and Beth Sundheim. Message understanding conference - 6: A brief history. In *Proceedings of the 16th Conference on Computational Linguistics, Copenhagen, Denmark*, pages 466–471, 1996.

Barbara J. Grosz and Luke Hunsberger. The dynamics of intentions in collaborative intentionality. *Cognitive Systems Research*, 7(2–3):259–272, 2006.

Barbara J. Grosz and Sarit Kraus. Collaborative plans for complex group action. *Artificial Intelligence*, 86(2):269–357, 1996.

Barbara J. Grosz and Sarit Kraus. The evolution of SharedPlans. In A. Rao and M. Wooldridge, editors, *Foundations and Theories of Rational Agencies*, pages 227–262. Springer, 1999.

Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

Barbara J. Grosz, Aravind Joshi, and Scott Weinstein. Centering: A framework for modelling the local coherence of discourse. *Computational Linguistics*, 21(2):203–225, 1995.

Claire Grover and Richard Tobin. Rule-based chunking and reusability. In *Proceedings of the Language Resources and Evaluation Confrerence*, 2006.

Claire Grover, Ben Hachey, and Chris Korycinski. Summarising legal texts: Sentential tense and argumentative roles. In Dragomir Radev and Simone Teufel, editors, *Proceedings of the NAACL-HLT Workshop on Automatic Summarization*, pages 33–40. Association for Computational Linguistics, 2003.

Alexander Gruenstein, John Niekrasz, and Matthew Purver. Meeting structure annotation: data and tools. In *Proceedings of the 6th SIGdial Workshop on Discourse and Dialogue*, September 2005.

Yufan Guo, Anna Korhonen, Maria Liakata, Ilona Silins, Lin Sun, and Ulla Stenius. Identifying the information structure of scientific abstracts: An investigation of three different schemes. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*, 2010.

Surabhi Gupta, John Niekrasz, Matthew Purver, and Daniel Jurafsky. Resolving "you" in multi-party dialog. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 227–230, 2007a.

Surabhi Gupta, Matthew Purver, and Daniel Jurafsky. Disambiguating between generic and referential "you" in dialog. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2007b.

Sangyun Hahn, Richard Ladner, and Mari Ostendorf. Agreement/disagreement classification: Exploiting unlabeled data using contrast classifiers. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 53–56. Association for Computational Linguistics, 2006.

Michael A. K. Halliday and Ruqayia Hasan. *Cohesion in English*. Longman, 1976.

William F. Hanks. *Referential practice: language and lived space among the Maya*. The University of Chicago Press, 1990.

William F. Hanks. Explorations in the deictic field. *Current Anthropology*, 46(2):191–220, 2005.

Ruqaiya Hasan. Situation and the definition of genres. In Allen D. Grimshaw, editor, *What's going on here? Complementary studies of professional talk (Volume Two of the Multiple Analysis Project)*, volume 43 of *Advances in Discourse Processes*. Ablex, 1991.

Marti A. Hearst. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

Marti A. Hearst. Automated discovery of wordnet relations. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

Peter A. Heeman. Speech actions and mental states in task-oriented dialogues. In *Proceedings of the AAAI Spring Symposium on Reasoning about Mental States*, 1993.

Dustin Hillard, Mari Ostendorf, and Elizabeth Shriberg. Detection of agreement vs. disagreement in meetings: Training with unlabeled data. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2003.

Kenji Hirohata, Naoaki Okazaki, Sophia Ananiadou, and Mitsuru Ishizuka. Identifying sections in scientific abstracts using conditional random fields. In *Proceedings of the International Joint Conference on Natural Language Processingc*, 2008.

Makoto Hirohata, Yousue Shinnaka, Koji Iwano, and Sadaoki Furui. Sentence extraction-based presentation summarization techniques and evaluation metrics. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2005.

Julia Hirschberg and Diane Litman. Empirical studies on the disambiguation of cue phrases. *Computational Linguistics*, 19(3):501–530, 1993.

Jerry Hobbs and David Andreoff Evans. Conversation as planned behavior. *Cognitive Science*, 4:349–377, 1980.

Chiori Hori, Sadaoki Furui, Rob Malkin, Hua Yu, and Alex Waibel. A statistical approach to automatic speech summarization. *EURASIP Journal of Applied Signal Processing*, 2003: 128–139, 2003a.

Takaaki Hori, Chiori Hori, and Yasuhiro Minami. Speech summarization using weighted finite-state transducers. In *Proceedings of the Interspeech Conference*, 2003b.

Chia-Hsin Hsieh, Chien-Lin Huang, and Chung-Hsien Wu. Spoken document summarization using topic-related corpus and semantic dependency grammar. In *Proceedings of the 2004 International Symposium on Chinese Spoken Language Processing*, pages 333 – 336, 2004.

Pei-Yun Hsueh. Audio-based unsupervised segmentation of meeting dialogue. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 2008a.

Pei-Yun Hsueh. *Meeting Decision Detection: Multimodal Information Fusion for Multi-Party Dialogue Understanding*. PhD thesis, School of Informatics, University of Edinburgh, 2008b.

Pei-Yun Hsueh and Johanna Moore. What decisions have you made?: Automatic decision detection in meeting conversations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2007a.

Pei-Yun Hsueh and Johanna D. Moore. Combining multiple knowledge sources for dialogue segmentation in multimedia archives. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1016–1023, 2007b.

Pei-Yun Hsueh and Johanna D. Moore. Automatic decision detection in meeting speech. In *Machine Learning for Multimodal Interaction, 4th International Workshop, MLMI 2007, Brno, Czech Republic, June 28-30, 2007, Revised Selected Papers*, pages 168–179. 2008.

Chien-Lin Huang, Chia-Hsin Hsieh, and Chung-Hsien Wu. Spoken document summarization using acoustic, prosodic and semantic information. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, page 4 pp., 2005.

Dell Hymes. *Foundations in Sociolinguistics: An Ethnographic Approach*. University of Pennsylvania Press, 1974.

Akira Inoue, Takayoshi Mikami, and Yoichi Yamashita. Improvement of speech summarization using prosodic information. In *Proceedings of Speech Prosody*, 2004.

Alejandro Jaimes, Kengo Omura, Takeshi Nagamine, and Kazutaka Hirata. Memory cues for meeting video retrieval. In *Proceedings of the 1st ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, 2004.

Adam Janin, Jeremy Ang, Sonali Bhagat, Rajdip Dhillon, Jane Edwards, Javier Marcías-Guarasa, Nelson Morgan, Barbara Peskin, Elizabeth Shriberg, Andreas Stolcke, Chuck Wooters, and Britta Wrede. The ICSI meeting corpus. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 364–367, 2003.

N. Jovanovic, H.J.A. op den Akker, and A. Nijholt. A corpus for studying addressing behaviour in multi-party dialogues. *Language Resources and Evaluation*, 40(1):5–23, 2006.

Natasa Jovanovic. *To Whom It May Concern - Addressee Identification in Face-to-Face Meetings*. PhD thesis, University of Twente, 2007.

Dan Jurafsky. Pragmatics and computational linguistics. In *Handbook of Pragmatics*, pages 578–604. Blackwell, 2005.

Daniel Jurafsky, Elizabeth Shriberg, and Debra Biasca. Switchboard swbd-damsl shallow-discourse-function annotation coders manual, draft 13. Technical Report 97-02, University of Colorado, Boulder. Institute of Cognitive Science, 1997.

Megumi Kameyama and Iaso Arima. Coping with aboutness complexity in information extraction from spoken dialogues. In *Proceedings of the International Conference on Speech and Language Processing*, 1994.

Megumi Kameyama, Goh Kawai, and Isao Arima. A real time system for summarizing human human spontaneous spoken dialogues. In *Proceedings of the International Conference on Speech and Language Processing*, 1996.

Brett Kessler, Geoffrey Numberg, and Hinrich Schütze. Automatic detection of text genre. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1997.

Dan Klein and Christopher D. Manning. Fast exact inference with a factored model for natural language parsing. In *NIPS 15*, 2002.

Thomas Kleinbauer, Stephanie Becker, and Tilman Becker. Combining multiple information layers for the automatic generation of indicative meeting abstracts. In *Proceedings of the European Workshop on Natural Language*, 2007.

Satoshi Kobayashi, Noriki Yoshikawa, and Seiichi Nakagawa. Extracting summarization of lectures based on linguistic surface and prosodic information. In *Proceedings of the ICSA/IEEE Workshop on Spontaneous Speech Procissing and Recognition*, 2003.

Balakrishna Kolluru, Heidi Christensen, and Yoshihiko Gotoh. Multi-stage compaction approach to broadcast news summarization. In *Proceedings of the Interspeech Conference*, 2005.

Natascha Korolija. Recycling cotext: The impact of prior conversation on the emergence of episodes in a multiparty radio talk show. *Discourse Processes*, 25(1):99–125, 1998a.

Natascha Korolija. *Episodes in talk: Constructing coherence in multiparty conversation*. PhD thesis, Linköping University, The Tema Institute, Department of Communications Studies, 1998b.

Natascha Korolija and Per Linell. Episodes: Coding and analysing coherence in multiparty conversation. *Linguistics*, 34(2):799–831, 1996.

Konstantinos Koumpis. *Automatic voicemail summarisation for mobile messaging*. PhD thesis, University of Sheffield, 2002.

Konstantinos Koumpis and Steve Renals. Automatic summarization of voicemail messages using lexical and prosodic features. *ACM Trans. Speech Lang. Process.*, 2(1):1, 2005.

Konstantinos Koumpis, Steve Renals, and M. Niranjan. Extractive summarization of voicemail using lexical and prosodic feature subset selection. In *Proceedings of the Eurospeech Conference*, 2001.

Klaus Krippendorff. *Content Analysis: An Introduction to Its Methodology*. Sage, 2004.

Rachel Laban. What are the human requirements for a technology that captures meeting information. Master's thesis, Department of Information Studies, University of Sheffield, 2004.

Sylvain Lamprier, Tassadit Amghar, Bernard Levrat, and Frederic Saubion. On evaluation methodologies for text segmentation algorithms. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*, pages 19–26, 2007.

Staffan Larsson. GoDiS 1.2 developers manual. Draft, 2000.

Beth Levin. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, 1993.

Lori Levin, Klaus Ries, Ann Thymé-Gobbel, and Alon Lavie. Tagging of speech acts and dialogue games in spanish call home. In *Proceedings of the ACL 1999 Workshop on Towards Standards and Tools for Discourse Tagging*, 1999.

Stephen C. Levinson. *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press, 1983.

Stephen C. Levinson. Activity types and language. In Paul Drew and John Heritage, editors, *Talk at Work: Interaction in Institutional Settings*, volume 8 of *Studies in interactional sociolinguistics*, pages 66–100. Cambridge Univiersity Press, 1992.

Elena Levy. *Communicating Thematic Structure in Narrative Discourse: The Use of Referring Terms and Gestures*. PhD thesis, University of Chicago, 1984.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Stan Szpakowicz Marie-Francine Moens, editor, *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81. Association for Computational Linguistics, 2004.

Shih-Hsiang Lin, Yi-Ting Chen, Hsin-Min Wang, and Berlin Chen. A comparative study of probabilistic ranking models for spoken document summarization. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages 5025 –5028, 2008.

Shih-Hsiang Lin, Yao-Ming Yeh, and Berlin Chen. Leveraging kullback-leibler divergence measures and information-rich cues for speech summarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):871–882, 2010.

Per Linell. *Approaching Dialogue: Talk, interaction and contexts in dialogical perspectives*, volume 3 of *IMPACT: Studies in Language and Society*. John Benjamins, 1998.

Linguistic Data Consortium. *TDT-4 Corpus Annotation Specification*, version 1.4 edition, 2002.

Agnes Lisowska. Multimodal interface design for the multimodal meeting domain: Preliminary indications from a query analysis study. Technical Report IM2.MDM-11, ISSCO, University of Geneva, November 2003.

Agnes Lisowska, Andrei Popescu-Belis, and Susan Armstrong. User query analysis for the specification and evaluation of a dialogue processing and retrieval system. In *Proceedings of the Language Resources and Evaluation Confrerence*, 2004.

Diane J. Litman and Rebecca J. Passonneau. Empirical evidence for intention-based discourse segmentation. In *Intentionality and Structure in Discourse Relations: Proceedings* *of a Workshop sponsored by the special interest group on generation of the association for computational linguistics*, pages 60–63, 1993.

Fei Liu, Feifan Liu, and Yang Liu. Automatic keyword extraction for the meeting corpus using supervised approach and bigram expansion. In *Proceedings of the IEEE Spoken Language Technologies Workshop*, 2008.

Feifan Liu and Yang Liu. Exploring correlation between rouge and human evaluation on meeting summaries. *Trans. Audio, Speech and Lang. Proc.*, 18(1):187–196, 2010.

Karen E. Lochbaum. A collaborative planning model of intentional structure. *Computational Linguistics*, 24(4):525–572, 1998.

Hans Peter Luhn. The automatic creation of literature abstracts. *IBM Journal of research and development*, 2(2):159 – 165, 1958.

Igor Malioutov and Regina Barzilay. Minimum cut model for spoken lecture segmentation. In *Proceedings of the International Conference on Computational Linguistics*, pages 25–32, 2006.

Inderjeet Mani, Gary Klein, David House, Lynette Hirschman, Therese Firmin, and Beth Sundheim. SUMMAC: A text summarisation evaluation. *Natural Language Engineering*, 8:43–68, 2002.

William Mann and Sandra Thompson. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281, 1988.

Daniel Marcu. *The Theory and Practice of Discourse Parsing and Summarization*. MIT Press, 2000.

Daniel Marcu and Laurie Gerber. An inquiry into the nature of multidocument abstracts, extracts, and their evaluation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2001.

Sameer Maskey and Julia Hirschberg. Automatic summarization of broadcast news using structural features. In *Proceedings of the Interspeech Conference*, 2003.

Sameer Maskey and Julia Hirschberg. Comparing lexical, acoustic/prosodic, discourse and structural features for speech summarization. In *Proceedings of the Interspeech Conference*, 2005.

Sameer Maskey and Julia Hirschberg. Summarizing speech without text using hidden markov models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2006.

Mark T. Maybury. Generating summaries from event data. *Information Processing and Management*, 31:735–751, 1995.

Iain McCowan, Daniel Gatica-Perez, Samy Bengio, Guillaume Lathoud, Mark Barnard, and Song Zhang. Automatic analysis of multimodal group actions in meetings. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):305–317, March 2005.

Kathleen McKeown, Regina Barzilay, David Evans, Vasileios Hatzivassiloglou, Judith L. Klavans, Ani Nenkova, Carl Sable, Barry Schiffman, and Sergey Sigelman. Tracking and summarizing news on a daily basis with columbia's newsblaster. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2002.

Kathleen McKeown, Julia Hirschberg, Michel Galley, and Sameer Maskey. From text summarization to speech summarization. In *Proceedings of the 2005 ICASSP Special Session on Human Language Technology: Applications and Challenges for Speech Processing.*, 2005.

Ivana Mikic, Kohsia Huang, and Mohan Trivedi. Activity monitoring and summarization for an intelligent meeting room. In *Proceedings of the Workshop on Human Motion*, 2000.

Marc Moens and Mark Steedman. Temporal ontology and temporal reference. *Computational Linguistics*, 14(2):15–28, 1988.

William Morgan, Pi-Chuan Chang, Surabhi Gupta, and Jason M. Brenier. Automatically detecting action items in audio meeting recordings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 96–103. Association for Computational Linguistics, 2006.

Jane Morris and Graeme Hirst. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1), 1991.

Megan Moser and Johanna D. Moore. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419, 1996.

Peter Mühlhäusler and Rom Harré. *Pronouns and People: The Linguistic Construction of Social and Personal Identity*. Blackwell, 1990.

Gabriel Murray. *Using Speech-Specific Characteristics for Automatic Speech Summarization*. PhD thesis, University of Edinburgh, 2008.

Gabriel Murray and Steve Renals. Meta comments for summarizing meeting speech. In A. Popescu-Belis and R. Stiefelhagen, editors, *Proceedings of the Workshop on Machine Learning and Multimodal Interaction*, pages 236–247. 2008.

Gabriel Murray, Steve Renals, and Jonathan Carletta. Extractive summarization of meeting recordings. In *Proceedings of the Interspeech Conference*, September 2005a.

Gabriel Murray, Steve Renals, Jonathan Carletta, and Johanna D. Moore. Evaluating automatic summaries of meeting recordings. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, June 2005b.

Gabriel Murray, Steve Renals, Johanna D. Moore, and Jean Carletta. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 367–374, 2006.

Gabriel Murray, Thomas Kleinbauer, Peter Poller, Thomas Becker, Steve Renals, and Jonathan Kilgour. Extrinsic summarization evaluation: A decision audit task. *ACM Transactions on Speech and Language Processing*, 6, 2009.

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. Interpretation and transformation for abstracting conversations. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–902. Association for Computational Linguistics, 2010a.

Gabriel Murray, Giuseppe Carenini, and Raymond Ng. Generating and validating abstracts of meeting conversations: a user study. In *Proceedings of the International Conference on Natural Language Generation*, 2010b.

John Niekrasz and Johanna Moore. Participant subjectivity and involvement as a basis for discourse segmentation. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 54–61, 2009.

Malvina Nissim, Shipra Dingare, Jean Carletta, and Mark Steedman. An annotation scheme for information status in dialogue. In *Proceedings of the Language Resources and Evaluation Confrerence*, 2004.

NIST. The Topic Detection and Tracking Phase 2 (TDT2) Evaluation Plan, Version 3.7, 1998.

NIST. *Automatic Content Extraction 2008 Evaluation Plan*, version 1.2d edition, 2008.

Andrew Olney and Zhiqiang Cai. An orthonormal basis for topic segmentation in tutorial dialogue. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 971–978, 2005.

David S. Pallett. The role of the national institute of standards and technology in darpa's broadcast news continuous speech recognition research program. *Speech Communication*, 37(1-2):3 – 14, 2002.

Vincenzo Pallotta, Violeta Seretan, Marita Ailomaa, Hatem Ghorbel, and Martin Rajman. Query types for meeting information systems: assessing the role of argumentative structure in answering questions on meeting discussion records. In *Proceedings of the 2006 COMMA Workshop on Modelling Meetings, Argumentation and Discourse (MMAD)*, 2006.

Vincenzo Pallotta, Violeta Seretan, and Marita Ailomaa. User requirements analysis for meeting information retrieval based on query elicitation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 1008–1015. Association for Computational Linguistics, 2007.

Rebecca J. Passonneau. Instructions for applying discourse reference annotation for multiple applications (DRAMA). 1997.

Rebecca J. Passonneau and Diane J. Litman. Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139, 1997.

Marie-Paule Péry-Woodley and Donia Scott. Computational approaches to discourse and document processing. *Traitement Automatique des Langues*, 47(2):7–19, 2006.

Lev Pevzner and Marti A. Hearst. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36, 2002.

Massimo Poesio. *The GNOME Annotation Scheme Manual, Version 4*, 2000.

Massimo Poesio. The MATE/GNOME scheme for anaphoric annotation, revisited. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, 2004a.

Massimo Poesio. Discourse annotation and semantic annotation in the gnome corpus. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics Workshop on Discourse Annotation*, 2004b.

Massimo Poesio and Ron Artstein. Anaphoric annotation in the ARRAU corpus. In *Proceedings of the Language Resources and Evaluation Confrerence*, 2008.

Massimo Poesio and Mijail A. Kabadjov. A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation. In *Proceedings of the Language Resources and Evaluation Confrerence*, 2004.

Andrei Popescu-Belis. Dimensionality of dialogue act tagsets. *Language Resources and Evaluation*, 42:99–107, 2008.

Wilfried Post, Erwin Elling, Anita Cremers, and Wessel Kraaij. Experimental comparison of multimodal meeting browsers. In *Proceedings of HCI International*, 2007.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The penn discourse treebank 2.0. In *Proceedings of the Language Resources and Evaluation Confrerence*, 2008.

Mark A. Przybocki, Jonathan G. Fiscus, John S. Garofolo, and David S. Pallett. 1998 hub-4 information extraction evaluation. In *Proceedings of the 1999 DARPA Broadcast News Workshop*, 1999.

Matthew Purver, Patrick Ehlen, and John Niekrasz. Shallow discourse structure for action item detection. In *Proceedings of the HLT-NAACL Workshop: Analyzing Conversations in Text and Speech (ACTS)*, pages 31–34, 2006a.

Matthew Purver, Konrad Körding, Thomas Griffiths, and Joshua Tenenbaum. Unsupervised topic modelling for multi-party spoken discourse. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 17–24, 2006b.

Matthew Purver, John Dowding, John Niekrasz, Patrick Ehlen, Sharareh Noorbaloochi, and Stanley Peters. Detecting and summarizing action items in multi-party dialogue. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, September 2007.

Matthew Purver, Raquel Fernández, Matthew Frampton, and Stanley Peters. Cascaded lexicalised classifiers for second-person reference resolution. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 306–309, 2009.

Gisela Redeker. Discourse markers as attentional cues at discourse transitions. In Kerstin Fischer, editor, *Approaches to Discourse Particles*, pages 339–358. Elsevier, 2006.

Rachel Reichman. Conversational coherency. *Cognitive Science*, 2(4):283–327, 1978.

Rachel Reichman. *Getting computers to talk like you and me*. MIT Press, 1985.

Norbert Reithinger, Michael Kipp, Ralf Engel, and Jan Alexandersson. Summarizing multilingual spoken negotiation dialogues. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2000.

Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. A keyphrase based approach to interactive meeting summarization. In *Proceedings of the IEEE Spoken Language Technologies Workshop*, 2008.

Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. Long story short - global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, 52(10):801 – 815, 2010.

Rutger Rienks and Dirk Heylen. Argument diagramming of meeting conversations. In *Proceedings of the ICMI Workshop on Multimodal Multiparty Meeting Processing*, pages 85–92, 2005.

Klaus Ries. Towards the detection and description of textual meaning indicators in spontaneous conversations. In *Proceedings of Eurospeech*, pages 1415–1418, 1999.

Klaus Ries. Segmenting conversations by topic, initiative and style. In *Proceedings of the 2001 SIGIR Workshop on Information Retrieval Techniques for Speech Applications*, 2001a.

Klaus Ries. *Accessing Spoken Interaction Through Dialogue Processing*. PhD thesis, Karlsruhe University, 2001b.

Klaus Ries and Alex Waibel. Activity detection for information access to oral communication. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2001.

Jan Robin Rohlicek, D. Ayuso, M. Bates, R. Bobrow, A. Boulanger, H. Gish, P. Jeanrenaud, M. Meteer, and M. Siu. Gisting conversational speech. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, 1992.

Andrew Rosenberg and Julia Hirschberg. Story segmentation of broadcast news in English, Mandarin, and Arabic. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 125–128, 2006.

Jennifer Rowley. *Abstracting and Indexing*. Clive Bingley, 1982.

Harvey Sacks. *Lectures on Conversation, Volume II*. Blackwell, 1992.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735, 1974.

Emanuel A. Schegloff. What type of interaction is it to be. In *Proceedings of the 18th Annual Meeting of the Association for Computational Linguistics*, pages 81–82. Association for Computational Linguistics, 1980.

Emanuel A. Schegloff. Reflections on quantification in the study of conversation. *Research on Language and Social Interaction*, 26:99–128, 1993.

Deborah Schiffrin. Meta-talk: Organizational and evaluative brackets in discourse. *Sociological Inquiry*, 50(3–4):199–236, 1980. Database Name: CSA Linguistics and Language Behavior Abstracts.

Deborah Schiffrin. Conversational coherence: The role of well. *Language*, 61(3):640–667, 1985. Database Name: CSA Linguistics and Language Behavior Abstracts.

Karin Kipper Schuler. *VerbNet: A broad-coverage, comprehensive verb lexicon*. PhD thesis, University of Pennsylvania, 2005.

Tanja Schultz, Alex Waibel, Michael Bett, Florian Metze, Yue Pan, Klaus Ries, Thomas Schaaf, Hagen Soltau, Martin Westphal, Hua Yu, , and Klaus Zechner. The isl meeting room system. In *Proceedings of the Workshop on Hands-Free Speech Communication (HSC-2001)*, 2001.

John R. Searle. *Speech acts: an essay in the philosophy of language*. Cambridge University Press, 1969.

John R. Searle. The classification of illocutionary acts. *Language in Society*, 5:1–24, 1976.

Melissa Sherman and Yang Liu. Using hidden Markov models for topic segmentation of meeting transcripts. In *Proceedings of the IEEE Spoken Language Technologies Workshop*, pages 185–188, 2008.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. The ICSI Meeting Recorder Dialog Act (MRDA) Corpus. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, pages 97–100, 2004.

John McH. Sinclair and R. M. Coulthard. *Towards an analysis of discourse: the English used by teachers and pupils*. Oxford University Press, 1975.

Carlota S. Smith. *Modes of Discourse*. Cambdrige University Press, 2003.

Karen Spärck Jones. What might be in a summary? In *Proceedings of Infomation Retrieval*, 1993.

Karen Spärck Jones. Reflections on TREC. *Information Processing & Management*, 31(3): 291 – 314, 1995.

Karen Spärck Jones. Patterns in the mind: language and human nature : Ray jackendoff. *Information Processing & Management*, 32(5):639 – 640, 1996.

Karen Spärck Jones. Automatic summarizing: Factors and directions. In Inderjeet Mani and Mark T. Maybury, editors, *Advances in Automatic Text Summarization*. MIT Press, 1999.

Karen Spärck Jones. Factorial summary evaluation. In *Proceedings of the Document Understanding Conference (DUC)*, 2001.

Karen Spärck Jones. Automatic summarising: The state of the art. *Information Processing & Management*, 43(6):1449 – 1481, 2007. Text Summarization.

Efstathios Stamatatos, Nikos Fakotakis, and George Kokkinakis. Text genre detection using common word frequencies. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2000.

Matthew Stone. Intention, interpretation and the computational structure of language. *Cognitive Science*, 28:781–809, 2004.

Matthew Stone and Alex Lascarides. Coherence and rationality in grounding. In *Proceedings of Workshop on the Semantics and Pragmatics of Dialogue*, 2010.

Stephanie Strassel, Mark Przybocki, Kay Peterson, Zhiyi Song, and Kazuaki Maeda1. Linguistic resources and evaluation techniques for evaluation of cross-document automatic content extraction. In *Proceedings of the Language Resources and Evaluation Confrerence*, 2008.

John Swales. *Genre Analysis*. Cambridge University Press, 1990.

Maite Taboada and William C. Mann. Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies*, 8(3):423–459, 2006.

Simone Teufel and Marc Moens. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4), 2002.

Shingo Togashi, Masaru Yamaguchi, and Seiichi Nakagawa. Summarization of spoken lectures based on linguistic surface and prosodic information. In *Proceedings of the IEEE Spoken Language Technology Workshop*, pages 34 –37, 2006.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 252–259, 2003.

David Traum. *A Computational Theory of Grounding in Natural Language Conversation*. PhD thesis, University of Rochester, 1994.

David Traum. Issues in multi-party dialogues. In F. Dignum, editor, *Advances in Agent Communication*, pages 201–211. Springer-Verlag, 2004.

David Traum and Elizabeth A. Hinkelman. Conversation acts in task-oriented spoken dialogue. *Computational Intelligence,* 8(3):575–599, 1992.

Simon Tucker and Steve Whittaker. Reviewing multimedia meeting records: Current approaches. In *Proceedings of the 2005 (ICMI) International Workshop on Multimodal Multiparty Meeting Processing*, 2005.

Simon Tucker and Steve Whittaker. Accessing multimodal meeting data: Systems, problems and possibilities. In *Lecture Notes in Computer Science: Machine Learning for Multimodal Interaction*, pages 1–11. 2007.

Simon Tucker and Steve Whittaker. Have a say over what you see: Evaluating interactive compression techniques. In *Proceedings of the International Conference on Intelligent User Interfaces*, 2009.

Simon Tucker, Steve Whittaker, and Rachel Laban. Identifying user requirements for novel interaction capture. In *Proceedings of the 5th International Conference on Measuring Behaviour*, 2005.

Masao Utiyama and Hitoshi Isahara. A statistical model for domain-independent text segmentation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 499–506, 2001.

Robin Valenza, Tony Robinson, Marianne Hickey, and Roger Tucker. Summarisation of spoken audio through information extraction. In *Proceedings of the ESCA ETRW Workshop on Accessing Information in Spoken Audio*, 1999.

Daan Verbree, Rutger Rienks, and Dirk Heylen. First steps towards the automatic construction of argument-diagrams from real discussions. In *Proceedings of the 1st International Conference on Computational Models of Argument, September 11 2006, Frontiers in Artificial Intelligence and Applications,* volume 144, pages 183–194. IOS press, 2006.

Wolfgang Wahlster, editor. *Verbmobil: Foundations of Speech-to-Speech Translation*. Springer, 2000.

Alex Waibel, Michael Bett, Micahel Finke, and Rainer Stiefelhagen. Meeting browser: Tracking and summarizing meetings. In *Proceedings of the DARPA Broadcast News Workshop*, 1998.

Katie Wales. *Personal pronouns in present-day English*. Cambridge University Press, 1996.

Marilyn Walker and Rebecca J. Passonneau. DATE: A dialogue act tagging scheme for evaluation of spoken dialogue systems. In *Proceedings of the First International Conference on Human Language Technology Research*, 2001.

Bonnie Webber. Genre distinctions for discourse in the penn discourse treebank. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2009.

Pierre Wellner, Mike Flynn, Simon Tucker, and Steve Whittaker. A meeting browser evaluation test. In *Proceedings of the 2005 Conference on Human Factors in Computing Systems (CHI)*, 2005.

Steve Whittaker, Rachel Laban, and Simon Tucker. Analysing meeting records: An ethnographic study and technological implications. In *Proceedings of the Workshop on Machine Learning and Multimodal Interaction*, pages 101–113. 2006.

Steve Whittaker, Simon Tucker, Kumutha Swampillai, and Rachel Laban. Deisgn and evaluation of systems to support interaction capture and retrieval. *Personal and Ubiquitous Computing*, 12(3):197–221, 2008.

Janyce M. Wiebe. Tracking point of view in narrative. *Computational Linguistics*, 20(2): 233–287, 1994.

Janyce M. Wiebe. References in narrative text. In Judy Duchan, Gail Bruder, and Lynne Hewitt, editors, *Deixis in Narrative: A Cognitive Science Perspective*, pages 263–286. 1995.

Theresa Wilson. Annotating subjective content in meetings. In *Proceedings of the Language Resources and Evaluation Confrerence*, 2008.

Maria Wolters and Mathias Kirsten. Exploring the use of linguistic features in domain and genre classification. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, 1999.

Stanton E. F. Wortham. Mapping participant deictics: A technique for discovering speakers' footing. *Journal of Pragmatics*, 25:331–348, 1996.

Chung-Hsien Wu, Chien-Lin Huang, and Chia-Hsin Hsieh. Spoken document summarization and retrieval for wireless application. In *Proceedings of the 2005 International Con-*

ference on Wireless Networks, Communications and Mobile Computing, volume 2, pages 1388 – 1393 vol.2, 2005.

Shasha Xie and Yang Liu. Improving supervised learning for meeting summarization using sampling and regression. *Computer Speech and Language*, 24(3):495–514, 2010.

Shasha Xie, Benoit Favre, Dilek Hakkani-Tür, and Yang Liu. Leveraging sentence wieghts in a concept-based optimization framework for meeting summarization. In *Proceedings of the Interspeech Conference*, 2009.

Weiqun Xu, Jean Carletta, Jonathan Kilgour, and Vasilis Karaiskos. *Coding Instructions for Topic Segmentation of the AMI Meeting Corpus*, 2005.

David M. Zajic, Bonnie J. Dorr, and Jimmy Lin. Single-document and multi-document summarization techniques for email threads using sentence compression. *Information Processing & Management*, 44(4):1600 – 1610, 2008.

Klaus Zechner. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001a.

Klaus Zechner. *Automatic Summarization of Spoken Dialogues in Unrestricted Domains*. PhD thesis, Carnegie Mellon University, 2001b.

Klaus Zechner. Automatic summarization of open-domain multiparty dialogues in diverse genres. *Computational Linguistics*, 28(4):447–485, 2002.

Klaus Zechner and Alon Lavie. Increasing the coherence of spoken dialogue summaries by cross-speaker information linking. In *Proceedings of the NAACL-01 Workshop on Automatic Summarization*, 2001.

Justin Jian Zhang and Pascale Fung. Speech summarization without lexical features for mandarin broadcast news. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2007.

Justin Jian Zhang, Ho Yin Chan ad Pascale Fung, and Lu Cuo. A comparative study on speech summarization of broadcast news and lecture speech. In *Proceedings of the Interspeech Conference*, 2007.

Justin Jian Zhang, Ho Yin Chan, and Pascale Fung. Extractive speech summarization using shallow rhetorical structure modeling. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1147 –1157, 2010.

Xiaodan Zhu and Gerald Penn. Comparing the roles of textual, acoustic and spoken-language features on spontaneous-conversation summarization. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 197–200. Association for Computational Linguistics, 2006.

Index