DISCOURSE PROCESSES, *35*(1), 1–31

Discourse Structure, Grounding, and Prosody in Task-Oriented Dialogue

Ilana Mushin
Department of Linguistics
University of Sydney, New South Wales, Australia

Lesley Stirling and Janet Fletcher
Department of Linguistics and Applied Linguistics
University of Melbourne, Victoria, Australia

Roger Wales
School of Humanities and Social Sciences
La Trobe University, Victoria, Australia

This article explores the relationship between discourse structure, grounding, and prosody in interactive discourse through an empirical analysis of task-oriented dialogue (the Australian Map Task corpus). Our focus is on the role that prosody plays in the process of *grounding*—the attainment and acknowledgment of mutual knowledge by discourse participants (Clark, 1996). We investigate how patterns of prosodic boundary strengths, pitch contour, pausing, and overlap relate to the structuring of "common ground units" (Nakatani & Traum, 1999), collaborative units that capture the points in the discourse at which common ground is established. We also examine the distribution of dialogue acts (Jurafsky, Shriberg, & Biasca, 1997) within common ground units to add to the emerging model of dialogue structure that takes into account the "joint action" feature of interactive discourse. Our results show that responses belonging to different types of common ground units have different prosodic profiles. These results have implications for computational and psychological modeling of dialogue structure as well as our understanding of the functions of prosody in interaction.

Correspondence and requests for reprints should be sent to Ilana Mushin, Department of Linguistics, University of Sydney, Sydney NSW, 2006, Australia. E-mail: ilana.mushin@linguistics.usyd.edu.au

This article explores the relationship between discourse structure and prosody as it relates to the process of *grounding* in dialogue. By prosody we mean patterns of fundamental frequency (F0) and timing that operate at the postlexical level in Australian English. In particular, we are interested in what aspects of the dialogue-specific discourse structures, those that represent the collaborative nature of interaction, exhibit prosodic regularities. This line of research has become increasingly important for the design and implementation of computational strategies for developing appropriate interactions between computers and people, apparent in the amount of recent activity in dialogue studies in computational linguistics (e.g., Cahn & Brennan, 1999; Dybkjaer, Hasida, & Traum, 2000), but it has also evolved out of a growing recognition that linguistic structure emerges from the interactive context in which language is used (e.g., Ochs, Schegloff, & Thompson, 1996).

Most empirical work examining prosodic signaling of discourse structure has focused on the use of acoustic cues to identify boundaries of discourse units in *monologue* (e.g., Grosz & Hirschberg, 1992; Hirschberg & Nakatani, 1996; Nakatani, Hirschberg, & Grosz, 1995; Swerts, 1997). These studies have found that a range of acoustic parameters associated with prosody, such as the position of pauses, final lengthening, and type of boundary tone, are good indicators of the boundaries between different discourse units. For example, Swerts (1997) found that pausing, pitch reset, and boundary tones were phonetic cues that contributed to the recognition of hierarchically organized discourse structure. Subjects used these cues to identify boundaries of discourse structure corresponding to paragraphs in written discourse.

More recently however, there has been an interest in examining how prosody may be used to signal discourse structure in dialogue. Auer (1996) and Wells and Peppé (1996), working in a conversation analysis framework, identified several prosodic factors (including rate of speech, loudness, and direction of pitch contour) that are associated with points of possible turn completion—the boundaries of turn construction units. Selting (1996) found that the patterns of "global pitch" and "loudness" were used by participants to distinguish between "activity types" (i.e., what participants were doing by speaking) in German conversations. Koiso, Horiuchi, Tutiya, Ichikawa, and Den (1998) found relationships between prosodic cues (duration, F0 contour, and relative pitch level), syntactic cues, and patterns of turn-taking and "backchannels." In the field of speech recognition research, Shriberg et al. (1998) showed that various prosodic cues (duration, F0, pause length, and speaking rate) were relevant for the automatic classification of dialogue "acts." Stirling, Fletcher, Mushin, and Wales (2001) similarly showed strong correspondences between the boundaries of dialogue acts and prosodic phenomena such as pitch reset and intonational phrase boundaries.

All of these studies take as their unit of analysis the "parts" of dialogue most akin to structural elements of monologic discourse. Dialogue acts (i.e., speech acts in dialogue), for example, can be analyzed as independent utterances by a single

speaker. Turn construction units, the minimal unit that can make up a turn of talk, share less with units of monologic discourse than dialogue acts do since their function is to set up points of potential speaker change in interaction. But turn construction units, as units of dialogue, also consist of single utterances by one speaker.

A fundamental aspect of interactive discourse that distinguishes it from monologue is that, in dialogue, the production of the discourse is a collaborative process between all participants. The type and form of one speaker's utterance is dependent on the type and form of another speaker's contribution to the dialogue. Different models of dialogue structure have labeled these collaborative structures "contributions" (Clark & Schaefer, 1987, 1989), "exchanges" (Sinclair & Coulthard, 1975), "games" (Carletta et al., 1996), or "adjacency pairs" (Sacks & Schegloff, 1973). Whatever the label, all of these models recognize that dialogue is not simply the juxtaposition of utterances but rather a reflection of the ongoing collaboration and negotiation of participants engaged in the act of communicating.

In terms of the relationship between prosody and discourse structure, it is clear that extra-segmental aspects of the speech signal—pitch, accent, timing, loudness, and so on—serve to delimit dialogue acts, and to some extent distinguish between them (as seen, for example, in Stirling et al., 2001). However we have yet to address the question of how prosody may be used to identify and distinguish dialogue structures that reflect the collaborative nature of dialogue. In particular, we ask the following question: Do the patterns of prosodic regularity that have been observed for dialogue acts actually reflect something of the collaboration between participants? If so, what aspects of the collaboration are reflected, and in what way does prosody signal these aspects?

These are important questions to address in the context of examining the role that prosodic phenomena play in the facilitation of communication. They contribute not only to our understanding of how human communication works, but also to the improvement of computational models of dialogue structure. These are the models that are driving the development of spoken language systems that cannot only adequately and creatively contribute to an unfolding discourse, but also utilize prosody in a way that seems natural.

This study reports on the first stages of this investigation—an empirical examination of naturally occurring (as opposed to experimentally elicited or computer-generated) spoken dialogues. We used a corpus of task-oriented dialogues taken from the Australian map task corpus, which is modeled on the HCRC map task corpus, to address these questions (Millar, Vonwiller, Harrington, & Dermody, 1994). By using task-oriented dialogues (see the Methodology section for details of the task), we remain in that domain of discourse that is most relevant to human–computer communication—communication for the purpose of achieving some specific task.

As mentioned earlier, there have been many different characterizations of collaborative structures, each based on a slightly different organizing principle. Here we

have chosen for our model of interactive discourse the "contribution" model developed in Clark and Schaefer (1987, 1989) and Clark (1996). The model takes *grounding*, the process by which information is collaboratively acknowledged as mutually shared by speech participants, as a basic principle of discourse organization. We present a summary of the model as we have applied it here in the next section.

We have chosen the contribution model partly because it has been usefully applied to the development of human–computer dialogue systems (see, for example, Benz, 2000; Brennan & Hulteen, 1995; Cahn & Brennan, 1999; Kreutel & Matheson, 2000). Grounding has also been proposed as the basis for identifying "units" of discourse structure that are used as the basic structures from which higher level intentional (i.e., goal-directed) structures are built (Core et al., 1999). These units, called "common ground units" (CGUs), are used here as the basic unit of collaborative structure. Details of CGU structure are given in the methodology section.

The results we present in this article concern whether three types of prosodic phenomena—boundary strength, final pitch contour, and interactional timing (interturn pause and overlap)—were good indicators of the degree of complexity of a CGU (collaborative unit). By "degree of complexity," we mean whether that collaboration involved just one contribution by each participant, or whether it involved more than one contribution before information could be said to have entered the common ground. The complexity of CGUs thus corresponds with expansions and repairs of various kinds as participants work toward the achievement of common ground. Here we hypothesized that the prosodic phenomena we investigated would exhibit regularities corresponding with different types of CGUs. Observation of such regularities would be taken as indirect evidence that prosody was being used by participants to signal the status of information regarding grounding, and to signal whether more contributions were required in order for information to be accepted as mutual knowledge.

We correlated each of the prosodic phenomena examined with CGUs of different complexities (i.e., how much difficulty the participants had establishing information as common ground). Extracts 1 and 2, from our corpus, are examples of what we call "simple" and more "complex" CGUs respectively.

(1) IF: Have you got a cross
IG: Yes
(2) IG: [what we have to actually] do is head is head west
IF: head west
IG: as in looking as if this is your map with north at the [top]
IF: [allright]

We hypothesized that both boundary strength and final contours, prosodic phenomena found to be relevant to the identification of both boundaries and types of dialogue acts, would display regularities according to the complexity of collaborative unit in these dialogues. We also hypothesized that timing of contributions rela-

tive to when the other participant is speaking would correlate with the degree of complexity of the collaborative unit.

Finally we examined the relationship between the CGUs and the dialogue acts from which they are composed. We compared dialogue act type with CGU complexity to determine the degree to which our prosodic results were simply epiphenomenal on how prosodic phenomena behave in dialogue acts.

GROUNDING AND DISCOURSE STRUCTURE

This section provides a summary of grounding, as developed by Clark and associates (e.g., Clark & Schaefer, 1987, 1989), and CGUs, as developed in Nakatani and Traum (1999). Here we discuss the basic principles that underlie this approach to discourse analysis, leaving the practicalities of categorization and annotation to our discussion of our methods.

Grounding takes place by virtue of a contribution being proposed by one participant and then evidence being given by the other that they have perceived (and understood) it. The evidence may be as minimal as "proceeding as usual," may consist of a head nod or other nonverbal cue, or may be an overt verbal acknowledgment or response.

Grounding is a dynamic process that can be modeled as a series of collaborative negotiations interspersed with moments of grounding. This leads to a change in the pragmatic and semantic status of the information considered to that point, as "grounded" pragmatic and semantic information is added to the participants' assumed common ground (cf. Poesio & Traum, 1997). So, though grounding is itself a cognitive notion, modeling the emergent nature of knowledge structures in discourse, it is not about the cognitive status of any individual mind (cf. Grosz, Joshi, & Weinstein, 1995, p. 5) "attentional states," which "model the discourse participants' focus of attention at any given point in the discourse." Rather, it is more concerned with the processes involved in acknowledging that information has been established as part of the common ground of the interlocutors.

Common ground can be established at several "levels" of action, as represented in Table 1, based on Clark (1996). In each case, the establishment of common ground is modeled as the collaborative effort of both participants. At level 1, the "prelinguistic, or nonlinguistic level," all that is required for grounding is that A executes some behavior (e.g., A points to a noticeboard) and B attends to that behavior (e.g., B looks at A pointing at the noticeboard). At level 2, A makes some linguistic gesture to which B responds in such a way as to acknowledge that a linguistic signal has been given (but not necessarily understood). Grounding at level 3 requires that the propositional content of A's contribution is acknowledged as being understood, whereas grounding at level 4 requires that there be some acknowledgment of the performance of a dialogue act by A. These levels are nested so that

TABLE 1
Levels of Action

Level	*Speaker A's Actions*	*Speaker B's Actions*	*Examples*
1	A is executing behaviour *t* for B	B is attending to behaviour *t* from A	A points to noticeboard; B looks at A pointing to noticeboard.
2	A is presenting signal *s* for B	B is identifying signal *s* from A	A makes some linguistic gesture to which B responds in such a way as to acknowledge that a linguistic signal has been given (but not necessarily understood).
3	A is signaling that *p* for B	B is recognizing that *p* from A	B acknowledges an understanding of the propositional content of A's contribution.
4	A is proposing joint project *w* to B	B is considering A's proposal of *w*	B acknowledges the performance of a dialogue act by A.

Note. From *Using Language* (p. 222), by H. H. Clark, 1996, Cambridge, England: Cambridge University Press. Copyright 1996 by Cambridge University Press. Reprinted with permission.

grounding at level 4 entails grounding at all of the other levels; grounding at level 3 entails grounding at levels 2 and 1, and so on.[1]

Clark's model of grounding has been proposed as the basis for the assignment of discourse chunks to units that reflect a basic level of collaborative structure (Traum, 1994, 1998). In particular, Nakatani and Traum (1999) have proposed CGUs as a unit of discourse structure on which higher levels that reflect intentional structure are based. CGUs are a formalization of Clark and Schaefer's (1989) "contributions," designed to allow consistent annotation of common ground status in dialogue. In Nakatani and Traum's framework, a CGU is considered to consist of all the linguistic material involved in achieving grounding of an initial contribution of information by one participant.

CGUs are units of discourse that represent the acknowledgment of common ground by both participants with respect to both parties understanding the propositional information of what was said—grounding at the "level of understanding," or level 3 grounding in Table 1. CGUs do not, for example, count the establishment of grounding at the level of perception, as in the example in Extract 2, repeated later, from our corpus, where IF (instruction-follower) signals recognition that IG (instruction-giver) has produced an utterance but queries his understanding of the content (possibly because he had been talking over IG at the start of her utterance, or because he thought he misheard what IG said). IG then adds

[1]The number and type of these levels is not completely fixed, and may depend on the nature of the communication involved. Brennan and Hulteen (1995), for example, identify seven levels of grounding that were required to develop appropriate kinds of feedback in a spoken language system. We have adopted Clark's (1996) classification because this was the classification used as the basis for identifying CGUs.

more information to clarify what was meant by her initial utterance. In this example, the first exchange established grounding at level of presentation (level 2) but the participants do not achieve grounding at the level of understanding (level 3) until the end of the fourth turn, where IF's response expresses an acknowledgment of the content of the instruction (and also that IG's contribution was an instruction). So all of the content in Extract 2 would be identified as a single complex CGU. We return to the discussion of the internal structure of CGUs in the following.

(2) IG: [what we have to actually]
do is head is head west
IF: head west
IG: as in looking as if this is your map with north at the [top]
IF: [allright]

Nakatani and Traum differentiate CGUs from other types of collaborative discourse structures, like dialogue games (Carletta et al., 1996) or adjacency pairs (Sacks & Schegloff, 1973). For example, unlike dialogue games, CGUs are not built from dialogue acts, or any other pragmatically determined structure. Rather, they are built from utterance tokens, established on the basis of syntax and prosody. In this way, CGUs have the potential to be considered as a basic unit of dialogue, bypassing analysis of the pragmatic goals of individuals in the interaction to capture a collaborative structure out of which units that represent the collaborative goals of the participants can be identified. They argue that CGUs capture only those parts of dialogue that relate to *mutual understanding*, whereas other approaches aim to capture all types of exchanges between participants.

METHODOLOGY

Our corpus consists of dialogues from the map task section of the Australian National Database of Spoken Language—ANDOSL (Millar, Vonwiller, Harrington, & Dermody, 1994). This corpus is closely modeled on the HCRC Map Task (Anderson et al., 1991). Participants worked in pairs, each with a map in front of them that the other could not see (although participants could see each other). One participant (the IG) had a route marked on their map and was required by the task to instruct the other (the IF) in drawing the correct route onto their own map. The maps were similar, but differed in the presence, location, and names of certain landmarks. Each pair of participants participated in two dialogues, swapping roles of instruction-giver and instruction-follower, and thus producing a first time and second time attempt at the task. Figures 1 and 2 provide examples of the maps used.

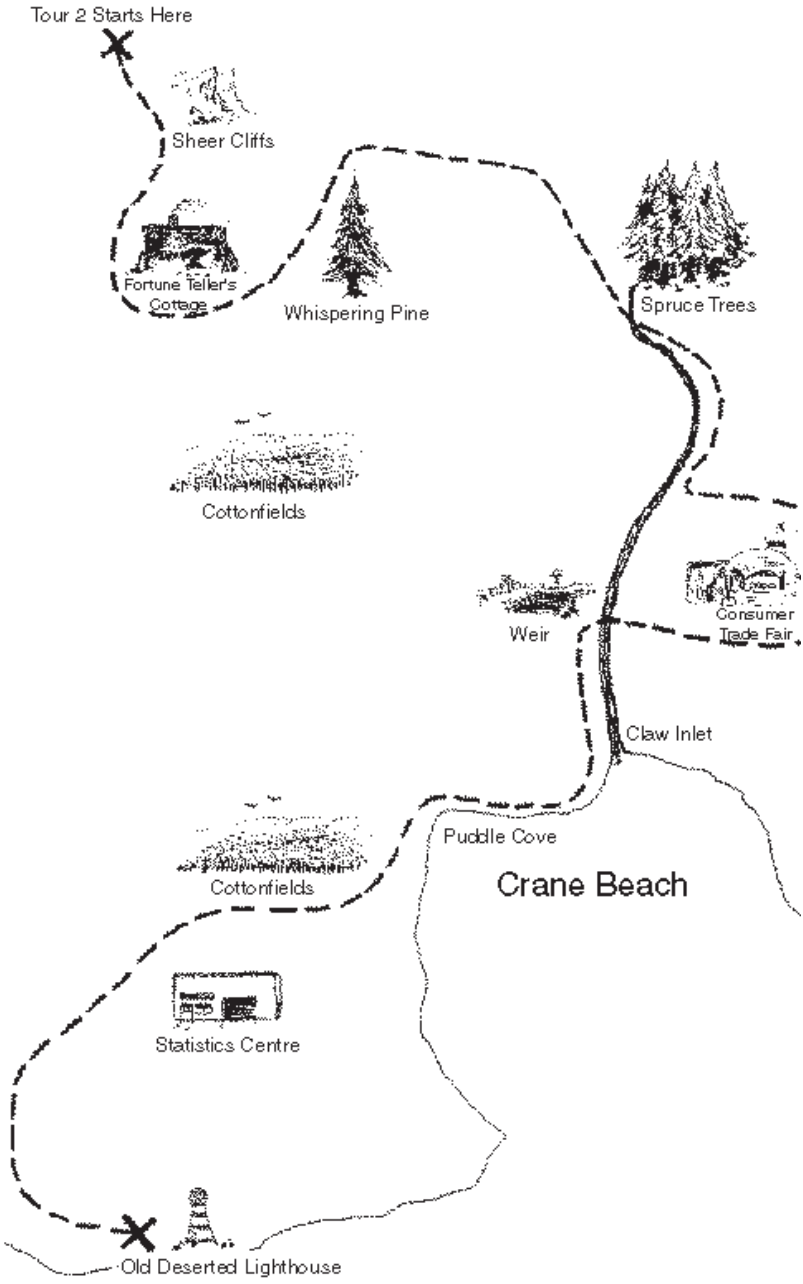

FIGURE 1 Map with route.

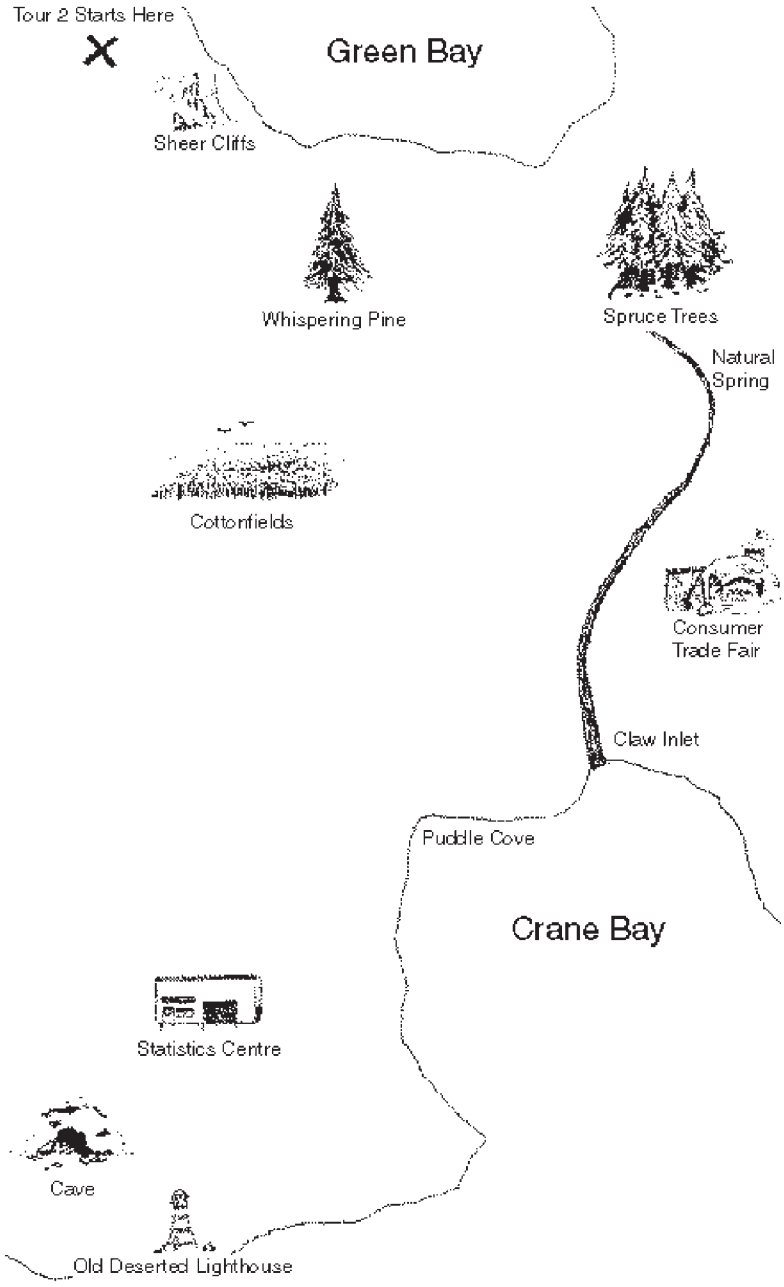

FIGURE 2 Map without route.

The four dialogues used here were from two pairs of participants: two dialogues from a pair "known" to one another (e.g., friends, colleagues), differing in which participant took the role of instruction-giver, and two dialogues from a pair "unknown" to one another, also differing in which participant took the role of instruction-giver. In each case, the pairs were mixed-sex. The dialogues were chosen randomly from the corpus and the participants were all speakers of General Australian English. The duration of the dialogues ranged between 485.93 and 810.24 s and their content totaled 7,393 words altogether. The prerecorded dialogues were copied from CD and digitized for analysis at 22 kHz using Entropic's ESPS/Waves + Speech analysis software running on a Sun workstation in the Phonetics Laboratory of the University of Melbourne. A complete orthographic transcription of the dialogues was carried out.[2]

Intonation and Timing Annotation

The dialogues were prosodically annotated for break indexes (degree of prosodic juncture) and phrase and boundary tones according to ToBI (Tone and Break Indexes) conventions using the Xwaves label function in ESPS/Waves+. Break index labels and tone labels were entered on separate label tiers. The ToBI prosodic transcription conventions for Australian English are detailed in Fletcher and Harrington (2001), and are closely modeled on the criteria developed for American English intonation (Beckman & Ayers Elam, 1994/1997). Break Indexes (BI) were labeled as follows: 4 (full intonation phrase boundary); 3 (intermediate intonation phrase boundary); 3p (disfluent intermediate intonation phrase boundary). All labels were aligned with the acoustic right edge of the last word of the relevant unit, using the acoustic waveform as a guide.

Boundary tones were labeled as follows: H–H% for a high rising tone (BI 4); L–L% for a falling or low tone (BI 4); L–H% for a low rising tone (BI 4); H–L% for a mid-level tone (BI 4); H for a high intermediate phrase boundary (BI 3); L for a low intermediate phrase boundary (BI 3). In all of these tone combinations, the *H* and *L* labels represent phonological high and low tones, respectively. The "–" after a tone label indicates that the tone is associated with an intermediate phrase boundary, and the "%" after a tone indicates that it is an intonational phrase-final boundary tone. The ToBI labeling was carried out by the third author of the article. It was conducted independently of discourse annotation.

Conventional interpretations of the four intonational phrase–final contour-types (i.e., H–H%, L–L%, L–H%, H–L%) are as follows: falling generally

[2]The small numbers of dialogues and speakers is partly due to the fact that these results report on the first stage of our research program, which is investigating the relationship between prosody and discourse structure using both naturally produced discourse and experimentally produced data. The labor intensiveness of the coding also precluded including more speakers at this stage of the research.

signals finality; high rising can either conclude a question or statement in Australian English (see Fletcher & Harrington, 2001, and Fletcher, Wales, Stirling, & Mushin, 2002, for closer examination of these differences); and low rising and mid-level usually form part of continuation contours, that is, signaling that there is more to come in the discourse. Intermediate phrase boundary tones also signal that there is more to come in the discourse.

The dialogues were also annotated on separate label tiers for the timing features of pause location and duration (in ms), and overlap location and duration (in ms).[3] These features were noted with respect to preceding and following talk (i.e., the utterance that immediately preceded the pause or overlap or immediately followed the pause or overlap).

These prosodic features of break index, phrase and boundary tones, and pause and overlap phenomena were selected to focus specifically on how certain intonational and timing phenomena interacted with CGU boundaries, and because related work in conversation analysis has demonstrated the importance of intonation (especially final contours) to being able to account for the meaning of acknowledgments in interaction (e.g., Gardner, 1998; Müller, 1996).

The pause and overlap labeling allowed us to extract information about the timing of each speaker contribution regarding the previous and following contributions to the talk. Regarding the preceding talk, contributions were analyzed as having a *pause before* the following contribution, *partially overlapping* with the preceding contribution, *completely overlapping*, *latching* with the preceding contribution (no pause or overlap), or *continued* by the same speaker without an intervening pause. Regarding the following talk, contributions were analyzed as having a *pause after* the preceding contribution, *partially overlapping* with the following contribution, *completely overlapping*, *latching* with the following contribution, or *continued* (if the next contribution was by the same speaker without an interceding pause).

Dialogue Act Annotation

Since a major goal of the project was precisely to investigate associations between prosodic characteristics and functional discourse categories, we coded both dialogue acts and CGUs independently from dividing the speech signal into prosodic units: For example, dialogue act coding did not presume prior division of the speech into intonational units (cf. the discussion in Zollo & Core, 1999). Coders used an unannotated orthographic transcription for the coding of discourse categories, and had no access to the speech signal. Dialogue acts were classified using both syntactic and pragmatic criteria. For example, an utterance of "yes" was clas-

[3]Overlap location and duration could be measured accurately because the dialogues were recorded on dual channels.

sified as an answer if it responded to a question; an acknowledgment if it responded to a statement; and an agreement if it responded to a directive.

The dialogues were initially coded for dialogue act using the "switchboard" version of DAMSL ("Dialogue Act Markup in Several Layers," Allen & Core, 1997). The DAMSL system was designed for the automatic annotation of dialogues according to types of dialogue acts. The original system had 220 labels based on both pragmatic and syntactic information. The "switchboard" version of DAMSL, developed in Jurafsky, Shriberg, and Biasca (1997) for a particular corpus of human–human telephone conversations, provides a simpler model that used 42 labels for annotation. Although we coded the original 42 labels in our full dialogue act annotations of the corpus, for the purposes of the analyses performed in this article we used the broadest classification labels of the switchboard system.[4] These are given in Table 2. In this system, the terms "forward-looking function" and "backward-looking function" are used for the broadest level of classification. As they are defined on syntactic–pragmatic grounds, "forward-looking function" and "backward-looking function" are theoretically independent from the sequential categories "initiation" and "response" that we used for classifying parts of CGUs. The nature of the relationship between these two types of classification are addressed in the Results section, where we examine the relationship between dialogue acts and CGUs.

The dialogue act coding was carried out by two researchers independently who then identified points of disagreement and resolved them to produce a consensual version for each dialogue.[5]

CGU Annotation

The dialogues were annotated for CGUs, following the procedures described in Nakatani and Traum (1999), by the same researchers who annotated the dialogue acts, also independently of the speech signal and prosodic annotation. CGUs consisted of all of the talk required to establish some information as common ground, independent of the delineation of dialogue acts (cf. conversational games, which are collaborative structures that are composed of dialogue acts—Carletta et al.,

[4]The SWBD-DAMSL coding scheme was selected in preference to the HCRC classification of moves developed specifically for the map task dialogues (Carletta et al., 1996) because of its finer granularity and focus on communicative functions of acts, rather than on a priori-determined sequence specific classifications. (See Stirling et al., 2001, for further details of the assessment of these two coding strategies for use in research into prosody and discourse structure.)

[5]Disagreements were few and mostly concerned the coding of "acknowledgment" and "agreement" categories. It turned out that the coders were using slightly different heuristics for determining when a contribution was an "acknowledgment" and when it was an "accept." The final version required an adjustment of the coding to reflect consistent heuristics for identifying different kinds of responses, as described earlier.

TABLE 2
Dialogue Act Categories

Forward-Looking Functions	*Backward-Looking Functions*
Statement	Agreement (accepts and rejects)
Influencing-addressee-future-action (questions and commands)	Understanding (acknowledgements)
Committing-speaker-future-action (offers and commits)	Answer
Other-forward-function (e.g., openings and closings)	

1996). The codes were entered on separate tiers in ESPS/Waves+, separated according to which speaker gave which contribution.

We called the first contribution to a CGU the "initiation phase" (analogous to Clark and Schaefer's "presentation phase"). Any further contributions to the CGU, by either participant, we called the "response phase" (analogous to Clark & Schaefer's "acceptance phase"). Initiation phases and response phases of CGUs were numbered and aligned with the speaker who began and ended each phase. Within each CGU we also coded the very first "response" utterance by the other participant, which typically established either that the initial information was entered into the common ground, or that further contributions were required in order to ground the information.

The most minimal CGUs are those that consist of an initiation contribution by one participant that is immediately acknowledged in the response phase as grounded minimally at the level of understanding (i.e., level 3 in Clark's model) by another participant. Extract 1, repeated in the following, is an example of a minimal, or simple, CGU, where IF asks a question that is answered by IG. IG's response acknowledges that she has understood both the fact that IF's utterance was a question and the content of that question.

(1) IF: Have you got a cross
IG: Yes

Nakatani and Traum (1999) also included in their characterization of CGUs contributions that were initiation phases that were not grounded in an adjacent response phase. These were either grounded at some point later in the dialogue ("discontinuous CGUs"), or were never grounded at the level of understanding ("abandoned CGUs"). The text corpus used for this study contained very few instances of discontinuous or abandoned CGUs, and so we have limited this study to the subset of CGUs that contained contributions that were adjacently grounded. That is, we only considered CGUs that contained fully grounded information at their close. We also only considered CGUs in which grounding was signaled vocally (e.g., by such vocal gestures as laughter) and omitted the few instances of exchanges in which grounding was presumed to have been signaled by nonvocal means (e.g., a gesture or facial expression).

Classification of CGUs

We have gone a step further than Nakatani and Traum (1999) in further classifying CGUs in terms of the internal characteristics of the grounding process—whether they consisted of a "simple" exchange of an initiating act and a responding act, as earlier, or whether their structure involved more contributions (by either speaker) before the CGU in question was considered to be complete (i.e., the information was acknowledged as entered into the common ground). The former type we called "simple" CGUs, whereas the latter type we called "complex" CGUs. Complex CGUs were further classified into four types based on pragmatic and sequential criteria, as follows:

1. Overlapping CGUs, where the grounding element of one unit was itself grounded with some verbal acknowledgment in the next CGU, as in Extract 3.

(3) IF: am I to the left-hand side or the right-hand side
of the d~
of your Galah Open-cut Mine
IG: **looking at it you're on the left** (finishes one CGU and starts another)
IF: okay

2. CGUs that contained further grounding elements by the speaker who made the initial contribution, as in Extract 4.

(4) IG: oh you've got Whispering Pine have you
IF: **yes**
IG: **right**

3. CGUs that contained more than one grounding element by the respondent, as in Extract 5.

(5) IG: that is uh as a point of looking at the Consumer Trade Affair
[right]
IF: **[okay] yeh**

4. CGUs that negotiated information at lower levels of grounding (e.g., the level of "presentation" or "locution" [Clark, 1996]), and were therefore considered part of a larger CGU that was grounded at the understanding level. For example, in Extract 6, the first response by IF functions to establish whether he has heard and understood the initial contribution by IG correctly. Only once this is established by IG's reiteration of the instruction does IF acknowledge that the initial contribution has been grounded at the level of understanding.

(6) IG: so you're] swee[ping east]
IF: [so am I sweep]ing right around [am I going east]

IG: [you're sweeping] east
IF: yeh okay well don't~~ yeh okay allright I'm going east

All four types of complex CGUs represent some kind of "expansion" of the canonical CGU exchange. In overlapping CGUs, the contribution that belongs to both CGUs functions to ground the first one and initiate the second; it is in turn grounded by a response to it. In multiple acknowledgment CGUs (i.e., types 2 and 3), the second response is not analyzed as "grounding," but is nevertheless some kind of further acknowledgment (e.g., by the initiator in type 2 or by the respondent in type 3). In the final case of complex CGUs, the first response is a signal that some part of the initial contribution had not entered into the common ground and that further collaboration was required.

The categories of complex CGUs were mostly, but not always, mutually exclusive. Overlapping CGUs (type 1) were not compatible with other types, but CGUs in some cases contained multiple acknowledgments by both speakers (types 2 and 3), or multiple acknowledgments in addition to further contributions at lower levels of grounding (type 4) before grounding was achieved at the level of understanding. For the purposes of this article we assigned each complex CGU to only one "primary" classification. In cases in which CGUs contained multiple acknowledgments by both speakers, they were split evenly between the two categories. We assigned CGUs that consisted of both multiple acknowledgments and further expansions to the "further contributions" category as this was the more complex of the two categories, in terms of the amount of information and interaction required to achieve grounding at the level of understanding.

These categories are not claimed to represent the only way that CGUs can be classified (cf. Core et al., 1999). Nevertheless they provide us with a useful basis for addressing differences in the role of prosody in CGUs.

Altogether there were 419 CGUs in the corpus. Of these, 241 were "simple" CGUs and 178 were "complex." Of the 178 complex CGUs, 47 (26%) were "overlapping" (type 1), 59 (33%) contained more than one acknowledgment by the same speaker (type 2), 52 (29%) contained multiple acknowledgments by the respondent (type 3), and 20 (11%) contained more complex negotiations of understanding until acknowledgment of mutual understanding of the initial dialogue act was reached (the level at which CGUs were delineated)—type 4.

We expected that because, according to our definition, complex CGUs differed from the simple type in terms of the structure of their response phase (i.e., all the contributions that respond to the initial contribution within a CGU), rather than in terms of their initiation phase, there should be no difference between simple and complex CGUs regarding the timing and prosodic profile of the initiation phase, nor with respect to the distribution of dialogue acts used to initiate new CGUs.

However, we predicted that there might be quantitatively recognizable properties of the first response contribution in complex CGUs that motivated their expan-

sion into further response contributions, and that these would be different from those found in simple CGUs (whose response phase consists of only one response contribution for grounding). We expected that these "recognizable" properties would involve a combination of grammatical, pragmatic, and prosodic properties. Here we focus on the extent to which simple and complex CGUs can be differentiated on prosodic grounds. We also looked at the extent to which simple and complex CGUs can be differentiated on the basis of dialogue act distribution in order to see whether the prosodic patterns we found were indeed symptomatic of the interactive process, or whether they were dependent on the types of dialogue acts expressed in either phase of the CGUs.

RESULTS

To check these hypotheses, data across all of the dialogues were compared for initiation and first response contributions in simple and complex CGUs for each of the four parameters identified above (break index, boundary tone, timing [before], and timing [after]), and for type of dialogue act of the initiation phase.

Our data set contains four speakers and four dialogues. The results presented here are the pooled results for all dialogues. Although there were clearly variations between dialogues, as one might expect in naturally occurring discourse (for example, the dialogues differed in length and in the amount of "extra task" chat), none of the results presented here is representative of any one particular dialogue or speaker.

The sparsity of data in some categories did have an effect on the way we have presented our results. Tables 3 through 8 show that the assumption of the chi-square statistic that the expected frequencies of each have a minimum value (conventionally 5, but see Everitt, 1977) cannot be sustained. When there is a cell whose expected values fall below the criterion of 5, the default computation is to add ½ to each cell to derive an approximation to the target statistic (or else it will not run). However in those cases where there is a lack of expected values in some cells, the reader should be alert to interpreting the inferential statistics with some caution.

The following labels have been used for the inferential statistical analysis: *correspondence* for the correspondences with break index, boundary tone, timing, and dialogue acts respectively; *pair* for individual dialogues that make up the corpus; *sequence* for the phases of the contribution (i.e., initiation and response phases); *type* for type of CGU (i.e., simple or complex).

Hierarchical loglinear analysis (computed in SPSS version 10) has been used in addition to the descriptive statistics for two reasons: The data are categorical, and the assumption of independence of the different speaker pairs needs to be tested. Loglinear analysis is one in which different models are tested for their fit to the

data. By definition, the "saturated" model, which includes all main effects and all their possible interactions, fits the data perfectly. The aim then is to find the "best" possible model that does not depart significantly from a good fit. The criterion of fit is expressed by chi-square values. "Best" can sometimes have a number of different interpretations since a number of models may all not depart significantly from a good fit. The choice between these alternatives is then made on the basis of which models have the smallest chi-square fit relative to other theoretical assumptions. The latter would include such issues as the number of factors in the model, the complexity of the interactions in it, and the relation between the model description and its interpretation in theoretical terms.

Because of the complexity of the designs used here, there are some cases where there are not enough free parameters in the model relative to the number of instances. This results in the model "overfitting" (demonstrated by zero likelihood ratio chi-square values). Though a "best model" is still computed, the analysis needs to be interpreted with caution. In such instances the results of chi-square analyses are reported. These have been computed as follows to deal with the assumption of independence. Where chi-square has been used on its own, the third variable has been treated as a "control" variable. That is, the pattern of frequencies is observed for one variable *independent* of a second variable for *one level* of the third variable, and then again for the *second level* of the third variable.

Grounding and Boundary Strength

Table 3 presents the results of our comparison of boundary strengths (measured in terms of ToBI) between simple and complex CGUs.

As expected, there was no significant difference between simple and complex CGUs regarding their initiation contribution break indexes. Examination of the first response contribution did show a significant difference between simple and complex CGUs with respect to break indexes. In particular, the first responses for simple CGUs had a high proportion of BI3 (corresponding to an intermediate phrase boundary) compared with complex CGUs and a relatively low proportion of BI4 (corresponding to a full intonation phrase)—22% of simple CGUs had BI3, compared with 10% of complex CGUs; only 67% of simple CGUs had BI4, compared with 82% of complex CGUs. That is, the first response contribution of a complex CGU was more likely to end with a full intonational phrase boundary (BI4) than the first response contribution of a simple CGU (see Discussion section for further details of this phenomenon).

The best model has generating class: pair*type, correspondence, sequence with a likelihood ratio (LR) $\chi^2 = 24.209$, $df = 52$, $p = 1.000$. What the "best" model signifies is that there is an interaction between the differences between the pairs and the CGU type, and the factors of correspondence and sequence are independent of any interaction. However this is a clear case of overfitting and the individual

TABLE 3
CGU Type × Sequence × Break Index (BI)

	BI4		*BI3*		*BI3p*		*No BI*		*Total*	
CGU Type	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Initiation phase										
Simple CGUs	227	94.2	7	2.9	5	2.1	2	0.8	241	100
Complex CGUs	165	92.7	4	2.3	9	5.0	0	0	178	100
Total	392	93.6	11	2.6	14	3.3	2	0.5	419	100
Response phase										
Simple CGUs	161	66.8	53	22.0	2	0.8	25	10.4	241	100
Complex CGUs	145	81.5	18	10.1	2	1.1	13	7.3	178	100
Total	306	73.0	71	16.9	4	1.0	38	9.1	419	100

Note. CGU = common ground units. $\chi^2(3, N = 419) = 2.81$, *ns* (initiation phase). $\chi^2(3, N = 419) = 12.69$, $p < 0.01$ (response phase).

chi-squares are $\chi^2(3, N = 419) = 2.81$, *ns* for initiations, and $\chi^2(3, N = 419) = 12.69$, $p < .01$ for responses.

Grounding and Boundary Tones

Table 4 presents the results of our comparison of boundary tones between simple and complex CGUs.

In contrast with the results for intonational phrase boundary strength reported earlier, there were significant differences found between simple and complex CGUs in relation to boundary tones in both initiation and response contributions. Regarding initiating contributions, a higher proportion of low falling (L–L%) boundary tones (42.1% vs. 29.5%) and a lower proportion of low rising (L–H%) boundary tones (18.5% vs. 29.9%) were found in the complex CGUs, compared with simple CGUs. Proportions of high rising (H–H%) boundary tones were not significantly different and there were proportionally few instances of other types of contours.

Like initiating contributions, the results for response contributions also show a higher proportion of low falling tones in complex CGUs than in simple CGUs (30.3% vs. 17.5%). This is an interesting result, as low falling tones typically are associated with completed statements and therefore might be expected to be more associated with the grounding of simple CGUs than complex CGUs. The relative prevalence of low falling tones in complex CGUs requires further scrutiny, especially in the context of its distribution with rising tones—a complex phenomenon in Australian English (see Fletcher et al., 2002, for a more detailed analysis of the role of rising tones in the Australian map task corpus). The results here suggest that the direction of pitch contour is not a good indicator of the complexity of a CGU.

TABLE 4
CGU Type × Sequence × Boundary Tone

	High Rise (H–H%)		*Low Fall (L–L%)*		*Low Rise (L–H%)*		*Mid-Level (H–L%)*		*High (H–)*		*Low (L–)*		*No Boundary*		*Total*	
CGU Type	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Initiation phase																
Simple CGUs	82	43.0	71	29.5	72	29.9	4	1.7	4	1.7	3	1.2	5	2.0	241	100
Complex CGUs	53	29.8	75	42.1	33	18.5	2	1.1	2	1.1	4	2	9	5.1	178	100
Total	135	32.2	146	34.9	105	25.1	6	1.4	6	1.4	7	1.7	14	3.3	419	100
Response phase																
Simple CGUs	51	21.2	42	17.5	55	22.8	13	5.4	21	8.7	28	11	31	12	241	100
Complex CGUs	40	22.5	53	30.3	39	21.9	13	7.3	7	3.9	12	6.8	13	7.3	178	100
Total	91	21.7	96	22.9	94	22.4	26	6.2	28	6.7	40	9.6	44	10	419	100

Note. CGU = common ground units. $\chi^2(6, N = 419) = 18.00$, $p < .01$ (initiation phase). $\chi^2(6, N = 419) = 17.23$, $p < .01$ (response phase).

The proportions of both high rising and low rising tones appears stable across the types of CGUs, however. Responses had a higher proportion of nonfinal (continuing) boundary tones (H–L%, H–, L–, and no boundary tone) than initiating contributions (32.5% vs. 7.8%). The higher proportion of "continuing" intonation contours in responses reflects the fact that responses were often followed by further talk by the same speaker who made the response (see Discussion section for further details of this phenomenon).

The best model has generating class: pair*type, correspondence*pair, correspondence*sequence with a LR $\chi^2 = 29.72$, $df = 73$, $p = 1.000$. That is, the best model has three two-way interactions. This is another clear case of overfitting and the individual chi-square values are $\chi^2(6, N = 419) = 18.00$, $p < .01$ (initiations) and $\chi^2(6, N = 419) = 17.23$, $p < .01$ (responses).

Correspondences With Timing

Timing relative to preceding contribution. Table 5 presents the results of our comparison of the timing of the initiation and response contributions with respect to the preceding contribution (by same or other speaker) for simple and complex CGUs.

Like the results for break indexes, the results for the timing of units with respect to immediately prior talk showed no significant differences between simple and complex CGUs for initiation phases, but did show significant differences with respect to first responses. That is, the results suggest that aspects of the timing of response contributions in relation to initiating contributions (the contribution preceding the response is the initiating contribution by definition) is an indicator of how much collaborative work is going to be required before grounding is established at the level of understanding.

In particular, there was a relatively high proportion of responses in complex CGUs whose onset occurred while the other speaker was still talking, resulting in a partial overlap (18.5% vs. 7.5%). There was no difference in the amount of completely overlapped response contributions between simple and complex CGUs. This means that, though these response contributions were timed to come in before the initiating contribution was completed, they did not come in so early as to be completely overlapped by the initiating contribution. We consider this result further in the discussion section.

The "best" model has generating class of two interactions: correspondence*pair, correspondence*type with a ratio LR $\chi^2 = 22.32$, $df = 22$, $p = .441$.

An almost equivalent "best" model is two interactions and one independent effect: correspondence*pair, correspondence*type, sequence.

TABLE 5
CGU Type × Sequence × Timing With Respect to Preceding Contribution

	Pause Before		*Overlap Before*		*Complete Overlap*		*Latch Before*		*Continued*		*Total*	
CGU Type	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Initiation phase												
Simple CGUs	98	40.7	38	15.8	1	0.4	44	18.2	60	24.9	241	100
Complex CGUs	62	34.8	34	19.1	3	1.7	27	15.2	52	29.2	178	100
Total	160	38.2	72	17.2	4	1.0	71	16.9	112	26.7	419	100
Response phase												
Simple CGUs	100	41.5	18	7.5	27	11.2	94	39.0	2	0.8	241	100
Complex CGUs	67	37.7	33	18.5	18	10.1	60	33.7	0	0	178	100
Total	167	39.8	51	12.2	45	10.7	154	36.8	2	0.5	419	100

Note. CGU = common ground units. $\chi^2(4, N = 419) = 4.59$, *ns* (initiation phase). $\chi^2(4, N = 419) = 13.06$, $p < .01$ (response phase).

Timing relative to next contribution. Table 6 presents the results of our comparison of timing of initiation and response contributions with respect to the next contribution (by same or other speaker).

The results for the timing of the unit following the target contribution also showed no significant difference for initiations between simple and complex CGUs, whereas the responses did show a difference of distribution. Contrary to the results for timing regarding preceding contribution, there seems to be a higher proportion of partially overlapped contributions in simple CGUs than in complex CGUs (7.5% vs. 3.9%), although the numbers are probably too small to interpret properly here.

More significantly, there were more latched contributions, ones where the second speaker timed his or her next turn to follow from the target contribution without any pause or overlap, in complex CGUs than in simple ones (22.5% vs. 12%). This suggests a smoother transition between speakers from first response to further contributions within the CGU and may be an indication that the contribution following the first response is still part of the same CGU.

The "best" model has generating class containing two, three-way interactions: correspondence*data*sequence, correspondence*sequence*type with a LR χ^2 = 11.11, df = 12, p = .519

Grounding and Dialogue Acts

Table 7 presents the distributions of forward-looking and backward-looking functions for simple and complex CGUs.

As predicted, the results for comparison of dialogue act distribution indicate no significant difference between simple and complex CGUs for initiations. However there is a significant difference in the distribution of first response contributions in simple CGUs from their distribution in complex CGUs. The clearest difference lies in the much higher proportion of forward-looking functions functioning as first responses (grounding elements) in complex CGUs (22% vs. 6%), a result predicted under the hypothesis that forward-looking functions are more likely to elicit further responses before grounding is fully achieved.

The best model is the saturated model: sequence*type*correspondence with a LR χ^2 = .000, df = 0, p = 1.000. This means that none of the effects is independent of the other. The individual chi-square values are $\chi^2(1, N = 419) = 0.75$, *ns* (initiations) and $\chi^2(1, N = 419) = 23.70$, $p < .001$ (responses).

We examined the differences between simple and complex CGUs regarding dialogue act distribution in first responses more closely. We subcategorized the backward-looking functions into the three main subcategories, as listed in Table 2: *agreement* functions (e.g., acceptances of requests), *understanding* functions (e.g., acknowledgments), and *answer* functions (e.g., answers to questions). Table 8 summarizes these results.

TABLE 6
CGU Type × Sequence × Timing With Respect to Following Contribution

	Pause After		*Overlap After*		*Complete Overlap*		*Latch After*		*Continued*		*Total*	
CGU Type	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Initiation phase												
Simple CGUs	111	46.1	31	12.9	2	0.8	90	37.3	7	2.9	241	100
Complex CGUs	69	38.8	40	22.5	2	1.1	60	33.7	7	3.9	178	100
Total	180	43.0	71	16.9	4	1.0	150	35.8	14	3.3	419	100
Response phase												
Simple CGUs	55	22.8	18	7.5	36	14.9	29	12	103	42.8	241	100
Complex CGUs	40	22.5	7	3.9	21	11.8	40	22.5	70	39.3	178	100
Total	95	22.7	25	5.9	57	13.6	69	16.5	173	41.3	419	100

Note. CGU = common ground units. $\chi^2(4, N = 419) = 7.64$, *ns* (initiation phase). $\chi^2(4, N = 419) = 9.95$, $p < .01$ (response phase).

TABLE 7
Distribution of Functions in the Initiation Phase

	Forward-Looking Functions		*Backward-Looking Functions*		*Total*	
CGU Type	*n*	*%*	*n*	*%*	*n*	*%*
Initiation phase						
Simple CGUs	210	88	31	12	241	100
Complex CGUs	160	90	18	10	178	100
Total	370	88	49	12	419	100
Response phase						
Simple CGUs	15	6	226	94	241	100
Complex CGUs	40	22	138	78	178	100
Total	55	13	364	87	419	100

Note. CGU = common ground units. $\chi^2(1, N = 419) = 0.75$, *ns* (initiation phase). $\chi^2(1, N = 419) = 23.70$, $p < .001$ (response phase).

TABLE 8
Distribution of Functions in the Response Phase

	Agreement		*Understanding*		*Answer*		*Forward-Looking Function*		*Total*	
CGU Type	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Response phase										
Simple CGUs	73	30	94	39	59	24	15	6	241	100
Complex CGUs	46	26	45	25	47	26	40	23	178	100
Total	119	28	139	34	106	25	55	13	419	100

Note. CGU = common ground units. $\chi^2(3, N = 419) = 27.48$, $p < .001$.

Table 8 shows that, in addition to a difference in the proportions of forward-looking functions, simple and complex CGUs also differed regarding the distribution of backward-looking functions. The first response parts of complex CGUs are fairly evenly distributed between acts of *agreement*, *understanding*, and *answers* to questions, whereas there is a slightly higher proportion of *understanding* functions in simple CGUs (39% vs. 30% *agreements* and 24% *answers*).

The results so far have pooled the different types of complex CGUs together. Recall that, although all complex CGUs involved an expansion of the basic initiation–response structure, there were differences in the ways that these expansions were manifested. We therefore recognized that there might be differences in dialogue act distribution between types of complex CGUs. Our final analysis was to separate the complex CGUs according to type and to compare their dialogue act distributions. This is illustrated in Table 9.

TABLE 9
Distribution of Complex CGUs and Dialogue Acts in the Response Phase

	Agreement		*Understanding*		*Answer*		*Total Backward-Looking Functions*		*Forward-Looking Functions*		*Total*	
Complex CGU Type	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Response phase												
Overlapping (type 1)	4	8	2	5	10	21	16	34	31	66	47	100
Other speaker expansion (type 2)	13	22	17	29	27	46	57	97	2	3	59	100
Same speaker expansion (type 3)	20	39	21	40	8	15	49	94	3	6	52	100
Grounding at other levels (type 4)	9	45	5	25	2	10	16	80	4	20	20	100
Total	46	25	45	25	47	26	13	71	40	29	178	100

Although these results are of a preliminary nature, they suggest that different kinds of complex CGUs do indeed manifest different profiles with respect to dialogue act distribution. For instance, there is a relatively high proportion of forward-looking functions in overlapped CGUs (where the unit that grounds the first CGU also initiates a new CGU)—78% of forward looking functions in responses were in overlapped CGUs. This result sets apart overlapped CGUs from other types of complex CGUs, which had very few instances of forward-looking functions as first responses, thereby accounting for the relatively high proportion of forward-looking functions for complex CGUs observed in the overall results.

DISCUSSION

In general, our results confirm that patterns of certain prosodic phenomena are consistent with a collaborative model of dialogue structure. Not only did initiation phases of CGUs overall behave differently to response contributions of CGUs with respect to prosody, but the responses themselves could also be differentiated based on whether they "finished" a CGU, or whether they were followed by other contributions before grounding was achieved.

The results show that, for the three prosodic parameters investigated (boundary strength, boundary tone, and timing of utterance), there was little or no variation according to the type of CGU in terms of the initiation phase. We also found no significant variation in the distribution of forward-looking functions and backward-looking functions in the initiation phase of both types of CGUs. These findings were consistent with our hypothesis that, if there were any measurable differences between these CGU types, they would be found in the first response contribution and not in the initiation phase.

Also consistent with our hypotheses were the results that the three prosodic parameters displayed differences between simple and complex CGUs in the first response contribution. Complex CGUs displayed more overlapped first responses, fewer incomplete intonation phrases (a break index of less than 4), and a higher proportion of low falling boundary tones (indicating completion) than the response contributions of simple CGUs.

Of the three prosodic parameters investigated, only boundary tones displayed differences between simple and complex CGUs for both initiation and response phases. Initiations of simple CGUs had a higher proportion of rising boundary tones (H–H% and L–H%) than those of complex CGUs. This was an unexpected result and one that requires further investigation. This result may reflect the type of dialogue act being performed in the initiation (e.g., questions vs. statements). However it might also be a result of individual speaker differences in preference for rising or falling contours in statement functions. The general Australian dialect of English is known for its high-rising final tones in nonquestion utterances, but

some speakers utilize them more than others. The patterns of rising tones as they related to discourse structure in our data have been addressed in Fletcher et al. (2002).

Overall, responses had a higher proportion of more minor types of boundary tones (including intermediate phrase boundaries and no boundary tone) than initiating contributions. This was unsurprising, since response phases often consisted of acknowledgment contributions (like *okay* or *yeh*), which indicated grounding at the understanding level, but were then followed by other talk by the same speaker continuing an intonational phrase. More specifically, the higher proportion of "nonfinal" contours (H–L%, L–, H–, and none) for simple CGUs compared with complex CGUs reflects a set of instances in which the respondent answers a yes–no question, and then makes a further response *in the same intonational phrase* that is new information that was itself grounded (and was therefore coded as a new CGU), as in Extract 7.

(7)	IG:	CGU1	Have you got a weir	(BI = 4, H–H%)
	IF:	CGU1	No	(–)
		CGU2	I've got this big natural spring	(BI = 4, L–L%)
	IG:	CGU2	No	(BI = 3, H–)
		CGU3	I don't have a natural spring okay	

These nonfinal contours typically also corresponded with break indexes of less than 4, accounting for the relatively high proportion of BI3s in responses in simple CGUs that was observed in our comparison of boundary strengths.

Regarding the timing parameter, there were interesting differences between the timing of a response contribution with respect to preceding and following contributions. The relatively high proportion of overlap in complex CGUs is due to the fact that, when the first response in a CGU is at least partially overlapping with the initiation contribution, it creates an environment in which both participants must work harder (i.e., make more contributions) in order for the acknowledgment of common groundedness to be clear. This suggests that it is the overlap that motivates the complexity of the CGU, and not the other way around. At least some of these overlaps resulted in the same speaker repeating an acknowledgment of grounding with no overlap, as in Extract 5, repeated below.

(5)	IG:	that is uh as a point of looking at the Consumer Trade Affair [right]
	IF:	**[okay] yeh**

In this example, IF's first acknowledgment response *okay* overlaps with the end of IG's statement initiating contribution (a *right* tag) and is immediately followed by another acknowledgment *yeh*. In this case, IF was indicating grounding at the level of understanding with her first response; however, as the first response was

overlapped, she reformulated the response to ensure that the acknowledgment of grounding had been mutually understood.

Regarding the timing of the contribution following the first response, the relatively high proportion of latching in complex CGUs is perhaps an indication of a smooth transition from first response to further grounding contributions within the CGU. It reflects the fact that, in complex CGUs, the next contribution after the first response utterance is still part of the same CGU. In simple CGUs, the next contribution after the first response contribution is a new CGU, which can be initiated by either speaker. Latching is perhaps a sign of both speakers working hard collaboratively to establish common ground in cases where common ground was not established in the first instance.

Whereas the distribution of dialogue acts in our corpus did correspond to a large extent with CGU structure (i.e., initiations with forward-looking functions and response phases with backward-looking functions), the distribution of dialogue acts in responses did show a dependence on the complexity of CGUs (e.g., there were more forward-looking functions in overlapping CGUs than in other kinds of CGUs). This indicates that, whereas there is a clear relationship between CGUs and dialogue acts, there are real differences between modeling dialogue as a sequence of speech acts and modeling it as a constant collaborative effort between participants toward the establishment of common ground. This strongly suggests that the patterns of prosodic regularity that we have observed in our data do indeed emerge from the collaborative effort. The results presented here suggest that prosody is doing more than just signaling individual speech acts (although it clearly does that too), as there are discernable patterns that distinguish responses that establish common ground in a simple exchange from those that consist of more complex sequences before common ground is established.

CONCLUSION

Most studies of prosody in dialogue have looked at relations between prosody and individual dialogue acts independent of their role in collaboration (Shriberg et al., 1998; Stirling et al., 2001). What we have shown here is that there is also a relationship between patterns of prosody and the ways that speakers work together toward mutual understanding. In some cases, the observed patterns clearly reflected some aspect of collaborative interaction, such as the role of overlap and latching in indicating degree of collaboration and common ground.

The results suggest that continued investigation of intonation and timing in the context of the "contribution" model has the potential to greatly enrich our understanding of the functions of prosody in interaction, especially as it reflects sequential complexities (Schegloff, 1992). In this article we have established that there are differences in the prosodic patterns found between simple and complex CGUs, mostly

without teasing apart the different kinds of complexity. Further investigation of the prosodic patterns found in each type of complex CGU is the next stage of the research. Such an analysis will provide us with a much finer grained picture of the relationship between intonation and timing phenomena and the types of contributions that motivate or anticipate more simple or complex collaborative structures.

Empirical investigation of the role that prosody plays in facilitating the achievement of mutual understanding is still in relative infancy. Yet such information is important for the development of computationally implementable models of dialogue that adequately reflect its collaborative nature. Here we have shown the utility of a corpus-based approach to the issue. We hope it will provide useful information for the development of systems that can better recognize patterns of complexity in the online negotiation of information, and so improve the manner in which humans and computers interact.

ACKNOWLEDGMENTS

Many thanks to Peter Kremer, Susan Brennan, and two reviewers for their helpful contributions.

REFERENCES

Allen, J., & Core, M. (1997). *Draft of DAMSL: Dialog Act Markup in Several Layers.* (Draft contribution for the Discourse Resource Initiative. Retrieved February 28, 2000 from http://www.cs.rochester.edu/trains/research/annotation)

Anderson, A., Bader, M., Bard, E., Boyle, E., Doherty, G., Garrod, S., et al. (1991). The HCRC map task corpus. *Language and Speech, 34,* 351–366.

Auer, P. (1996). On the prosody and syntax of turn-continuations. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: Interactional studies* (pp. 57–100). Cambridge, England: Cambridge University Press.

Beckman, M. E., & Ayers Elam, G. (1994/1997). *Guide to ToBI Labelling–Version 3.0.* (Electronic text and accompanying audio example files available at http://ling.ohiostate.edu/Phonetics/E_ToBI/etobi_homepage.html)

Benz, A. (2000). Chains and the common ground. *Gothenberg Papers in Computational Linguistics 00–5,* 181–184.

Brennan, S. E., & Hulteen, E. A. (1995). Interaction and feedback in a spoken language system: A theoretical framework. *Knowledge-Based Systems, 8,* 143–151.

Cahn, J. E., & Brennan, S. E. (1999). A psychological model of grounding and repair in dialog. *Proceedings of AAAI Fall Symposium on Psychological Models of Communication in Collaborative Systems,* 25–33.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., & Anderson, A. (1996). *HCRC dialogue structure coding manual* (HCRC/TR–82). Edinburgh, Scotland: University of Edinburgh, Human Communication Research Centre.

Clark, H. H. (1996). *Using language*. Cambridge, England: Cambridge University Press.

Clark, H. H., & Schaefer, E. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes, 2,* 1–23.

Clark, H. H., & Schaefer, E. (1989). Contributing to discourse. *Cognitive Science, 13,* 259–294.

Core, M., Ishizaki, M., Moore, J., Nakatani, C., Reithinger, N., Traum, D., et al. (1999). *The report of the Third Workshop of the Discourse Resource Initiative.* Chiba, Japan: Chiba University and Kazusa Academia Hall.

Dybkjaer, L., Hasida, K., & Traum, D. (Eds.). (2000). *Proceedings of the 1st SIGdial workshop on discourse and dialogue.* Association for Computational Linguistics.

Everitt, B. S. (1977). *The analysis of contingency tables.* London: Chapman & Hall.

Fletcher, J., & Harrington, J. (2001). High rising terminals and fall-rise tunes in Australian English. *Phonetica, 58,* 215–229.

Fletcher, J., Wales, R., Stirling, L., & Mushin I. (2002). A dialogue act analysis of rises in Australian English Map Task dialogues. *Proceedings of Speech Prosody 2002, Aix-en-Provence, France,* 299–302.

Gardner, R. (1998). Between listening and speaking: The vocalisation of understandings. *Applied Linguistics, 19,* 204–224.

Grosz, B., & Hirschberg, J. (1992). Some intonational characteristics of discourse structure. *Proceedings of the International Conference on Spoken Language Processing, Banff,* 429–432.

Grosz, B., Joshi, A., & Weinstein, S. (1995). Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics, 21,* 203–225.

Hirschberg, J., & Nakatani, C. (1996). A prosodic analysis of discourse segments in direction-giving monologues. *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz,* 286–293.

Jurafsky, D., Shriberg, L., & Biasca, D. (1997). *Switchboard SWBD–DAMSL shallow-discourse-function-annotation coder's manual, draft 13* (Tech. Rep. No. TR 97–02). Boulder: University of Colorado at Boulder, Institute for Cognitive Science. Retrieved February 28, 2000 from http://stripe.colorado.edu/~jurafsky/manual.august1.html

Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., & Den, Y. (1998). An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech, Special issue on Prosody and Conversation, 41,* 295–322.

Kreutel, J., & Matheson, C. (2000). Obligations, intentions and the notion of conversational games. *Gothenberg Papers in Computational Linguistics 00–5,* 107–114.

Millar, J., Vonwiller, J., Harrington, J., & Dermody, P. (1994). The Australian national database of spoken language. *Proceedings of the International Conference on Acoustics Speech and Signal Processing, 94,*197–200.

Müller, F. (1996). Affiliating and disaffiliating with continuers: Prosodic aspects of recipiency. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: Interactional studies* (pp. 131–176). Cambridge, England: Cambridge University Press.

Nakatani, C., Hirschberg, J., & Grosz, B. (1995). Discourse structure in spoken language: Studies on speech corpora. *Working notes of the AAAI–95 Spring Symposium on Empirical Methods in Discourse Interpretation* (pp. 106–112). Menlo Park, CA: American Association for Artificial Intelligence.

Nakatani, C., & Traum, D. (1999). *Coding discourse structure in dialogue (Version 1.0*) (Tech. Rep. No. UMIACS–TR–99–03). College Park: University of Maryland, Institute for Advanced Computer Studies.

Ochs, E., Schegloff, E. A., & Thompson, S. A. (Eds.). (1996). *Interaction and grammar.* Cambridge, England: Cambridge University Press.

Poesio, M., & Traum, D. (1997). Conversational acts and discourse situations. *Computational Intelligence, 13,* 309–347.

Sacks, H., & Schegloff, E. (1973). Opening up closings. *Semiotica, 7,* 289–327.

Schegloff, E. (1992). Repair after next turn: The last structurally provided defence of intersubjectivity in conversation. *American Journal of Sociology, 97,* 1295–1345.

Selting, M. (1996). Prosody as an activity-type distinctive cue in conversation: The case of the so-called "astonished" questions in repair initiation. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: Interactional studies* (pp. 231–270). Cambridge, England: Cambridge University Press.

Shriberg, E., Bates, R., Stolcke, A., Taylor, P., Jurafsky, D., Ries, K., et al. (1998). Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and Speech, Special issue on Prosody and Conversation, 41,* 443–492.

Sinclair, J. M., & Coulthard, R. M. (1975). *Towards an analysis of discourse: The English used by teachers and pupils.* Oxford, England: Oxford University Press.

Stirling, L., Fletcher, J., Mushin, I., & Wales, R. (2001). Representational issues in annotation: Using the Australian map task corpus to relate prosody and discourse structure. *Speech Communication, 33,* 113–134.

Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America, 101,* 514–521.

Traum, D. (1994). *A computational theory of grounding in natural language conversation.* Unpublished doctoral dissertation, Department of Computer Science, University of Rochester, NY. (Also available as TR 545, Department of Computer Science, University of Rochester)

Traum, D. (1998). *Notes on dialogue structure.* Unpublished manuscript. Retrieved February 28, 2000 from http://www.cs.umd.edu/users/traum/DSD/

Wells, B., & Peppé, S. (1996). Ending up in Ulster: prosody and turn-taking in English dialects. In E. Couper-Kuhlen & M. Selting (Eds.), *Prosody in conversation: Interactional studies* (pp. 101–130). Cambridge, England: Cambridge University Press.

Zollo, T., & Core. M. (1999). Automatically extracting grounding tags from BF tags. *Proceedings of the 37th Meeting of the Association for Computational Linguistics, Workshop on Standards and Tools for Discourse Tagging, College Park,* 109–114.