Utterance Units in Spoken Dialogue[*]

David R. Traum[1] and Peter A. Heeman[2]

[1] University of Maryland, College Park, MD 20742 USA
[2] University of Rochester, Rochester, NY 14627 USA

Abstract. In order to make spoken dialogue systems more sophisticated, designers need to better understand the conventions that people use in structuring their speech and in interacting with their fellow conversants. In particular, it is crucial to discriminate the basic building blocks of dialogue and how they affect the way people process language. Many researchers have proposed the *utterance unit* as the primary object of study, but defining exactly what this is has remained a difficult issue. To shed light on this question, we consider grounding behavior in dialogue, and examine co-occurrences between turn-initial grounding acts and utterance unit boundary signals that have been proposed in the literature, namely prosodic boundary tones and pauses. Preliminary results indicate high correlation between grounding and boundary tones, with a secondary correlation for longer pauses. We also consider some of the dialogue processing issues which are impacted by a definition of utterance unit.

1 Introduction

Current spoken dialogue systems tend to restrict the allowable interaction patterns to simple exchanges with the user. Unlike natural human conversation, turn-taking is generally fixed to the expression of a single sentence or speech act, with little flexibility about when the turn can change. To move beyond this will require a better understanding of the nature of spoken dialogue. An important starting point is a clear formulation of the basic units of language production and comprehension. Our opinion is that the best units of dialogue for an artificial system to attune to are the very same ones that humans use. Since spoken dialogue systems are meant to interact with humans, they must be able to use appropriate conventions in responding in a timely and appropriate fashion.

It has often been claimed that for spoken dialogue, *utterances* rather than *sentences* are the primary object of study [3, 4]. But just what *are* utterances? How are they built up from more primitive bits of speech, and how do they cohere with other utterances by both the same and other speakers? Following Bloomfield [2], the term *utterance* has often been vaguely defined as "an act of speech." The problem is that action comes in many different types and sizes. As discourse analysis of written text concerns the relationships between different sentences rather than sentence internal relationships, discourse analysis of spoken dialogue should concern the relationships between utterances. Finding an appropriate definition of *utterance units* is thus an important starting

point for distinguishing utterance-internal language processes (e.g., phonology, speech repairs) from those that operate at a discourse level, (e.g., turn-taking, grounding, rhetorical relations). We describe some of the needs for a precise utterance unit definition in Section 3.

Analysts have proposed many different definitions of utterances and *utterance units*. The *turn* is the unit of dialogue that has most often been proposed for study as a basic utterance unit. Fries [8], for example, uses the term *utterance unit* to denote those chunks of talk that are marked off by a shift of speaker. Other definitions are based on a variety of information, including syntactic (sentence, clause), semantic/pragmatic (a single proposition or speech act), or prosodic (tunes, stress, silence). We consider some of these in more detail in Section 2. The units we will evaluate specifically are spans of speech terminated by prosodic cues: boundary tones and pauses.

Evaluation of the relative efficacy of utterance units proposals is also difficult. Many of the uses for which one needs such a notion are internal cognitive processes, not directly observable in human conversation. One aspect which *is* observable is the way one participant in a conversation reacts to the speech of another. Much of this reaction will be responsive speech itself, especially if other channels, such as eye gaze, are not available (as in telephone dialogues). Such methods have been used previously in studies of turntaking (e.g., [6, 24]), using the turn-transition as a locus for observation of unit features.

In this study we look not just at the existence of a turn transition, but *whether* and *how* the new speaker reacts to what has been said. According to the view of conversation as a collaborative process, inspired by Clark [4], the conversants are not just making individual contributions, but working together to augment their common ground. This process of *grounding* involves specifically acknowledging the installments of others. By examining which speech is acknowledged, we can have some insight into which installments are viewed as successful, and by correlating these with proposed boundary signals, we can also evaluate the utility of the signals. In Section 4, we discuss the grounding phenomena and the coding scheme we used to mark relatedness, which captures the most immediate aspect of grounding – how a new speaker initially responds to prior speech.

In Section 5, we describe how we used this scheme, along with prosodic markings to label a dialogue corpus, to study this issue, with the results presented in Section 6. We then conclude with some observations on how to extend this study and some implications for spoken dialogue systems.

2   Proposals for Utterance Units

Although there is not a consensus as to what defines an utterance unit, most attempts make use of one or more of the following factors.

- Speech by a single speaker, speaking without interruption by speech of the other, constituting a single *Turn* (e.g. [7, 8, 22, 27].
- Has syntactic and/or semantic completion (e.g. [7, 22, 21, 32]).
- Defines a single speech act (e.g. [22, 16, 20]).

- Is an intonational phrase (e.g. [12, 9, 7, 11]).
- Separated by a pause (e.g. [22, 11, 29, 33]).

While the turn has the great advantage of having easily recognized boundaries,[3] there are several difficulties with treating it as a basic unit of spoken language. First of all, the turn is a multi-party achievement that is not under the control of any one conversant. Since the turn ends only when another conversant speaks, a speaker's turn will have only an indirect relation to any basic units of language production. If the new speaker starts earlier than expected, this may cut off the first speaker in midstream. Likewise, if the new speaker does not come in right away, the first speaker may produce several basic contributions (or units) within the span of a single turn.

From a purely functional point of view, many analysts have also found the turn too large a unit for convenient analysis. Fries, for example, noted that his utterance units could contain multiple sentences. Sinclair and Coulthard [31], found that their basic unit of interaction, the *exchange*, cut across individual turns. Instead, they use *moves* and *acts* as the basic single-speaker components of exchanges. A single turn might consist of several different moves, each of which might be part of different exchanges.

Sacks et. al, [27] present a theory of the organization of turns as composed of smaller *turn-constructional units* (TCUs). At the conclusion of each TCU there occurs a *transition-relevance place* (TRP), at which time it is appropriate for a new speaker to take over (or the current speaker may extend her turn with a subsequent TCU). TCUs thus form a more basic utterance unit, from which turns (and perhaps exchanges) can be built. TCUs have been posited to consist of differing types of syntactic contributions, including lexical, phrasal, clausal, and sentential constructions. Much subsequent work on turn-taking (*e.g.,* [6, 24, 7]) has tried to analyze what features are used to signal a TRP. The features that were examined included syntactic completions, pauses, and various prosodic features including boundary tones.

The difficulty with the syntactic, semantic, and pragmatic categories is that they can be very difficult to segment in spoken dialogue. For instance, should one pick the smallest or largest unit which applies? Very often, speakers add to their current installment, using such cue words as "and", "then", and "because" to link to what has come before. Moreover, speakers involved in spontaneous conversation do not always speak in well-formed units according to these criteria. Many actual productions are either fragmentary, ungrammatical, or both. Sometimes, as in the case of repaired speech, there are *no* correct units as produced, these only emerge from a (sometimes interactive) dialogue process. While it would be interesting to take another look at these candidates, more effort must first be spent on devising reliable rules for coding them consistently in spontaneous spoken dialogue.

We therefore focus here on the prosodic features of speech. When people speak, they tend not to speak in a monotone. Rather, the pitch of their voice, as well as other characteristics, such as speech rate and loudness, varies as they speak. (Pitch is also referred to as the fundamental frequency, or f_0 for short.) Pierrehumbert [25] presented a model of intonation patterns that later was taken as part of the basis for the ToBI (Tones

---

[3] Difficulties still remain, such as how to count turns when more than one conversant is speaking, and in determining whether a particular utterance counts as a backchannel item.

and Break Indices) annotation scheme [30]. This model describes English intonation as a series of highs (**H**) and lows (**L**). The lowest level of analysis deals with stressed words, which are marked with either a high or low *pitch accent*. The next level is the *intermediate phrase*, which consists of at least one stressed word, plus a high or low tone at the end of the phrase. This *phrasal tone* controls the pitch contour between the last pitch accent and the end of the phrase. The highest level of analysis is the *intonational phrase* (IP), which is made up of one or more intermediate phrases and ends with an additional high or low tone, the *boundary tone*, which controls how the pitch contour ends.

Another way in which a turn can be segmented is by pauses in the speech stream [29, 33]. Pause-delimited units are attractive since pauses can be detected automatically, and hence can be used to segment speech into chunks that are easier to process with a speech recognizer or parser. It is still controversial, however, whether pause-delimited speech is a good candidate for a definition of an utterance unit. For one thing, pauses can occur anywhere in the speaker's turn, even in the middle of a syntactic constituent. Also, oftentimes the speaker's pitch level will continue at the same level before and after the pause. There is also often some silence around the point of disfluency during a speech repair.

There have also been some more radical proposals for smaller basic units. For example, Poesio [26] proposed that each word or morpheme be a basic unit, accessible to all aspects of the discourse processing, including semantics, reference resolution, and dialogue management. This position has some attractions because there is psycholinguistic evidence that people do in fact use context even at such basic levels and are able to act appropriately (e.g., by gazing at an object) as soon as the necessary words to disambiguate the meaning are heard [34]. Additionally, this position allows a uniform style of processing and representation for phenomena like pronoun resolution which have both intra- and inter-sentential components.

3   Uses for Utterance Units

We feel, however, that such a step as word by word discourse processing goes too far — there is a legitimate role for sub-turn utterance units in distinguishing local from discourse phenomena in dialogue processing. While certain aspects of the local context will be necessary for an agent to attend to, many issues can be better dealt with by attending to the utterance unit and performing some kinds of processing only within UU boundaries, while performing others only between utterances. We briefly review some of these issues in this section.

3.1   Speech Repairs

Heeman and Allen [13] propose that speech repairs should be resolved very early in speech processing. Speech repairs are instances where speakers replace or repeat something they just said, as in this example from [15]:

we'll pick up   a tank of   uh   the tanker of oranges
            reparandum   *alteration*

In the above example the speaker intends that the text marked as the *reparandum* be replaced by the *alteration*. Speakers tend to give strong local clues both to mark the occurrence of speech repairs [18] and to help the hearer determine the intended correction [17].

One type of speech repair of particular interest is the *fresh start*, in which speakers abandon what they are saying, and start over. Exactly how much is abandoned is not explicitly signaled. Consider the following example from the TRAINS corpus [14] (d93-13.3 utt7), with intonation phrase endings marked as '**%**'.

> two hours **%**
> um okay **%**
> so um I guess I 'd like to take um
> and tankers only take orange juice right **%**

In this example a fresh start occurs starting on the fourth line. From the context, it seems that the speaker is canceling the text starting the word "okay" – this is just the current utterance, not material communicated in previous IPs.

Not all same-turn repairs are speech repairs. Consider the following example (d93-26.2 utt41):

> oh that wouldn't work apparently **%**
> wait wait **%**
> let's see **%**
> maybe it would **%**
> yeah it would **%**
> right **%**
> nineteen hours did you say **%**

In the above example the speaker changes what he/she is saying, yet it doesn't seem as if a speech repair is taking place. The repair is at a more interactional level, since the speaker needs the participation of the hearer to complete it. For true speech repairs, the processing is much more localized, and perhaps strictly to the current utterance unit. A firm definition for utterance units would thus provide a good metric for deciding which repairs could be done locally, and which repairs need more information and have a more global effect.

3.2   Speech Act Identification

While there is a variety of action that proceeds at different levels throughout spoken dialogue, determination of utterance unit sizes can be helpful for distinguishing certain types of action. As an example, the theory of Sinclair and Coulthard [31] is composed of hierarchical ranks of action, some of which occur across speaker turns, and some of which occur within the turn of a single speaker. The notion of an utterance unit could be useful in isolating which span of speech encapsulates a particular rank of action. Also, the multi-level conversation act theory proposed in [36] used an utterance unit distinction. In that proposal, turn-taking acts occurred within utterance units, grounding acts (see Section 4) occurred at the utterance unit level, and traditional core speech acts were composed of multiple utterance units.

Obviously, for such a theory, the nature of an utterance unit will affect how some acts are classified. For example, the set of grounding acts from [36] included acts for `repair` and `continue` functions. Trivially, smaller utterance units would lead to more continue acts, but consider the case in which a repair (e.g., a speech repair) occurs *within* an utterance unit. In this case, the function of the entire unit with respect to the previous context would be considered, and the unit might be seen as performing a continue function. If, however, such a unit were split up so that the alteration was in one unit and the reparandum is in a previous unit, the new unit might be termed a repair.

While this might not pose a serious problem for processing within a system, it does make it much harder to judge validity of such a coding scheme. Although there is a general agreement that it is important for dialogue systems to attend to action, there is still not much agreement on what the proper set of actions should be. For action types which are sensitive to an utterance unit distinction, performing any evaluation of the consistency of action marking becomes next to impossible if the proper set of units are not held constant for different coders.

3.3   Dialogue Management

There are a number of ways in which a good definition of utterance units would facilitate the dialogue management tasks of a spoken NL dialogue system. For one thing, the basic decision of when and how to respond to input from a user can be informed by awareness of utterance units. Even if full discourse processing of language proceeds within an utterance unit, as suggested by [26], the decision of when to respond still mainly follows regular patterns, occurring at regular places [27]. Awareness of the markers for turn-endings and TRPs can help with the decision of when to respond and when to wait for further input.

More importantly, the nature of the previous utterance unit can be a good signal of *how* to respond. Suppose, for instance, the speaker projects a desire to continue speaking. In this case, unless there is some major problem with the previous input, the system can issue a backchannel type response, without detailed processing (yet) of the implications of the speaker's contribution.

On the other hand, suppose there is some problem in understanding the current input. In this case, there are two main options left to the system: (1) wait for the user to say more and hopefully clarify or repair the problem, (2) notify the user of the problems, requesting some sort of repair. Attending to utterance unit production can help in deciding between these two options – in the midst of a unit, it will be preferable to wait for the user. With any luck, the speaker will decide to repair the problem, in which case the speech repair techniques proposed by [13] will fix the problem. At an utterance unit boundary, the dialogue manager can shift strategies, dependent on the anticipated continuation by the user. Given a signal that the user would like to continue her turn, if the problem is some sort of ambiguity or under-specification that is likely to be cleared up by future speech, the system can continue to wait. If, however, the problem is an error in understanding (or disagreement about) the previous utterance, the system can decide to break in with a repair.

An atunement to utterance units can also help a dialogue manager decide how to structure the speech output of the system. By using regular utterance units, a system

could provide the same resources to a user for providing feedback of understanding. Also, a system could structure the output in order to request help from the user, such as by ending an utterance unit at a point where the user could fill in a referring expression that the system is having difficulty calculating. In general, the division of contributions into more manageable sized utterance units, which can build up turns of speech, will be a more efficient mechanism for flexibly coordinating the contributions of the conversants to reach a mutual understanding of what has been said.

4 Grounding and Relatedness

As our method for checking the adequacy of proposed utterance unit boundaries, we consider the phenomenon of *grounding*: the process of adding to common ground between conversants [5]. Clark and Schaefer present a model of grounding in conversation, in which *contributions* are composed of two phases, *presentations* and *acceptances*. In the presentation phase, the first speaker specifies the content of his contribution and the partners try to register that content. In the acceptance phase, the contributor and partners try to reach the *grounding criterion*: "the contributor and the partners mutually believe that the partners have understood what the contributor meant to a criterion sufficient for the current purpose." Clark and Schaefer describe several different methods that are used by the partners to accept the presentation of the contributor. These include feedback words such as *ok*, *right*, and *mm-hm*, repetition of the previous content, and initiation of a next relevant contribution. Completions and repairs of presentations of the contributor also play a role in the grounding process.

Traum and Allen [35] built on this work, presenting a *speech acts* approach to grounding, in which utterances are seen as actions affecting the state of grounding of contributed material. In addition to acts which present new material, there are *acknowledgment* acts which signal that the current speaker has understood previous material presented by the other speaker, as well as *repairs* and *requests for repair*. Acknowledgment acts include three types, *explicit* acknowledgments which are one of the feedback words, whether they appeared as a backchannel or not, *paraphrases* of material presented by the other speaker, and *implicit* acknowledgments, which display understanding through conditional relevance. Explicit acknowledgments have also been studied by Novick and Sutton [23], who catalogued the interaction sequences that they play a part in.

4.1 Relatedness

For the purposes of determining an appropriate utterance unit boundary, we are not as concerned with whether something is ultimately grounded, but rather whether the responding agent *starts* the grounding process (with a repair or some component of the acceptance phase), and *which* previous utterances are being grounded. We thus use a less detailed labeling of utterances than that of [35], with a slightly different focus. We lump the repair, request for repair, and the paraphrase and implicit categories of acknowledgment together into one category we call *related*. While, for a particular utterance, it can be difficult to judge *which* one of of these functions is being performed, it is usually straightforward to determine whether or not it performs one of them. We

also separate out the *explicit* acknowledgments, since, while they generally perform some sort of acknowledgment, it is not possible to tell with certainty *what* they are acknowledging. Likewise, for utterances that follow backchannels and other turns that consist solely of these signals, there is no real content for the new turn to be related *to*. The third major category is for *unrelated* utterances, which either introduce new topics, or cohere only with previous speech by the same speaker and do not play a grounding role towards presentations by the other speaker. Our final category is for those utterances for which it is *uncertain* whether they are related or not.

4.2 Relatedness Coding Scheme

We used single character annotations to indicate relatedness of utterances. These characters are shown on the line above the text that they label. The coding scheme is explained below, including some examples.

Category: Explicit Ack
Label: e
Description: This category includes all utterances by a new speaker that begin with an **explicit** acknowledgment – the first word of the turn by the new speaker is a word like "okay", "right", "yeah", "well", or "mm-hm" (also "no" and "nope" in a negative polarity context). This category is also used for repair requests for which it is not clear exactly what should be repaired (e.g., an utterance consisting solely of "what?").
Example: d93-18.3 utt11-12

```
u: so we have to start in Avon
     e
s: okay
```

Category: Related
Description: This category is for those utterances that *demonstrate* a relationship to a previous utterance. This can be a demonstration style acknowledgment (i.e., repetition or paraphrase of a previous utterance by the other speaker), a repair or repair request, or some kind of continuation. Any second part to an *adjacency pair* [28] (such as an answer to a question, or the acceptance or rejection of a proposal) fits in this category. In addition, the recency of the relationship is marked as follows:
Subcategories:

Label	Description
0	utterances **related to** the **most recent** UU by the previous speaker.
1	**related** to UU one **previous** to most recent but *not* related to most recent.
2	related to UU two previous to most recent but nothing more recent.
etc.	higher numbers for related to utterances further back.
,	**related** to previous material by the other speaker, but it is **unclear** to the coder which unit they are related to.

Category: Unrelated

Label: u

Description: these utterances are **unrelated** to previous speech by the previous speaker. This could be either the introduction of a new topic, or could show a relation to previous speech by the same speaker, while ignoring intervening speech by the other speaker.

Example: d93-18.3 utt13-18

```
U: how long does it take to bring engine one to Dansville
   0
S: three hours
   e
U: okay <sil> and then <sil> back to Avon to get the bananas
   0
S: three more hours si(x)- six in all
   u
U: how long does it take to load the bananas
```

Category: Uncertain

Label: ?

Description: The coder is **uncertain** whether or not these utterances relate to previous speech by the other speaker.

Combination Category: After Explicit

Label: X-e

Description: A suffix of **-e** is used for those utterances which are **after** turns which consist *only* of **explicit** acknowledgments (or backchannels). Since these utterances do not have any content that new material *could* be related to, the categories for unrelated to last (**u**, **1**, etc.), are not very significant. Therefore, we append the marking **-e** to the other category in these cases.

Subcategories:

Label	Description
u-e	material which is unrelated to previous material by other speaker, but for which the last UU by the other speaker was an explicit acknowledgment.
1-e	material which is related to the penultimate utterance unit by the other speaker, but for which the last utterance unit by the other speaker contained just an explicit acknowledgment. This signal is often used for popping sub-dialogues, such as in the following example, in which the last turn by speaker **S** refers to the first utterance by speaker **U**.

Example: d93-18.3: utt51-54

```
U: how long does that take <sil> again
   0
S: that ta(ke)- <sil> so just to go from Bath to Corning
   e
U: mm-hm
   1-e
S: two hours
```

5 Data

As part of the TRAINS project [1], a corpus of problem-solving dialogs has been collected in which a railroad system manager (labeled U) collaborates with a planning assistant (labeled S) to accomplish some task involving manufacturing and shipping goods in a railroad freight system. Since this corpus contains dialogues in which the conversants work together in solving the task, it provides natural examples of dialogue usage, including a number of tough issues that computer systems will need to handle in order to engage in natural dialogue with a user. For instance, our corpus contains instances of overlapping speech, backchannel responses, and turn-taking: phenomena that do not occur in collections of single speaker utterances, such as ATIS [19]. For our current study, we looked at 26 dialogues from the TRAINS-93 corpus [14]. This totaled over 6000 seconds of spoken dialogue comprising 1366 total turn transitions.

5.1 Prosodic markings

All turn-initial utterances are marked with respect to their direct relatedness to the previous full or partial intonation phrase (IP) by the previous speaker. Full IP's are terminated by a boundary tone (labeled **%**). Partial phrases are anything from the last complete IP to the beginning of speech by the new speaker. If there was no previous IP by the previous speaker than the entire turn so far is counted as a partial IP. The amount of silence between turns is also noted, labeled below in seconds within square brackets (e.g., **[.42]**).

5.2 Relatedness Markings

Each turn-transition was marked as to how the initial installment of the new turn related to the last few completed or uncompleted IPs produced by the other speaker, using the coding scheme described in Section 4.2. For cases of overlapping speech, the speech of the new speaker was marked for how it relates to the current ongoing installment by the continuing speaker. For simultaneous starts of IPs, only the speech by the new speaker was marked for how it related to the previous speech by the current speaker. The continuing speech of one speaker was not marked for how it related to embedded speech by the other speaker, including true backchannels,

Table 1 summarizes this coding scheme described in Section 4.2, as used to mark turn-initial relatedness to utterance units from the previous turn.

5.3 Example

Below we show some examples of labeled dialogue fragments. The prosodic and relatedness markings are shown above the line they correspond to. The first example shows a simple sequence of **0** and **e** relations, all following boundary tones, with clean transitions. The first response by S shows a relationship to the first contribution by U. Then the last two start with explicit acknowledgments.

Label	Description
e	**explicit** acknowledgment (e.g., "okay", "right", "yeah", "well", or "mm-hm")
0	**related to** the **most recent** utterance by the previous speaker
1	**related** to the UU one **previous** to the most recent but *not* to the most recent
2	related to utterance two previous to the last one (and not to anything more recent)
,	**related** to previous material by the other speaker, but it is **unclear** to the coder whether they are related to the immediately previous UU (which would be marked **0**), or to an UU further back (which would be marked **1**, or **2**, etc.)
u	**unrelated** to previous speech by the old speaker
?	**uncertain** whether these utterances relate to previous speech by the other speaker
u-e 1-e	the same meaning for the first item, but follows a turn by the other speaker consisting only of an item marked **e**

Table 1. Relatedness Markings

Example: d93-13.2: utt18-22

```
                                         % [.42]
U: how long is it from Elmira to Dansville
   0                            % [1.23]
S: Elmira to Dansville is three hours
   e  %
U: okay um so why don't uh
                                      % [1.42]
   I send engine two with two boxcars to Corning
   e  %
S: okay
```

6 Results

6.1 Prevalence of Grounding Behavior

Tabulating the markings on the beginning of each turn yields the results shown in Table 2. This shows how the next utterance is related to what the other speaker has previously said, and so gives statistics about how much grounding is going on. Of all turns, 51% start with an explicit acknowledgment (category **e**); 29% are related to previous speech of the other speaker (one of the following categories: **0 1 2 , 1-e 2-e ,-e**); 15% are unrelated to previous speech of the other speaker, but follow an acknowledgment (**u-e**); 2% are possibly related or possibly unrelated (category **?**), and only 3% are clearly unrelated and do not follow an acknowledgment.

These results give strong evidence of grounding behavior at turn transitions. Fully 80% of utterance display grounding behavior, while another 15% occur in positions in which (according to the theory of [35, 36]) further grounding is unnecessary. It is only in 3-5% of turn transitions in which a lack of orientation to the contributions of the other speaker is displayed.

Category	#	%
Explicit	696	51%
Related	400	29%
Unrelated after Explicit	199	15%
Unrelated	42	3%
Uncertain	29	2%
Total	1366	100%

Table 2. Prevalence of Grounding Behavior

6.2 Boundary Tones

Table 3 shows how relatedness correlates with the presence of a boundary tone on the end of the preceding speech of the other speaker. Here, we have subdivided all of the markings into two groups, those that occur at a smooth transition between speaker turns (*clean transitions*), and those in which the subsequent speech overlaps the previous speech (*overlap*). For the overlap cases, we looked for a boundary tone on the last complete word before the overlapping speech occurred. The distribution of the overlaps into tone and no-tone categories is still somewhat problematic, due to the potential projectability of IP boundaries [10]: a new speaker may judge that the end of a unit is coming up and merely anticipate (perhaps incorrectly) the occurrence of a tone. Thus for some of the entries in the second to last column, there is a boundary tone which occurs after the onset of the new speaker's speech.

Type	Clean Transitions			Overlaps		
	Tone	No Tone	% Tone	Tone	No Tone	% Tone
e	501	24	95%	77	94	45%
0	267	17	95%	16	41	28%
1,2	7	4	64%	7	11	39%
,	9	4	69%	1	6	14%
1,2-e	7	0	100%	3	0	100%
u	18	7	72%	2	15	12%
u-e	186	2	99%	5	6	45%
?	17	3	85%	6	3	67%
Total	1012	61	94%	117	176	40%

Table 3. Boundary Tones and Relatedness

For the clean transitions, we see that more than 94% of them follow a boundary tone. Of more interest is the correlation between the relatedness markings and the presence of a boundary tone. For explicit acknowledgments and utterances that are related to the last utterance, we see that 95% of them follow a boundary tone. For transitions in

which the next utterance relates to an utterance prior to the last utterance, or is simply unrelated, we see that only 64% and 72% of them, respectively, follow a boundary tone. This difference between related to last (**0**) and related to prior and unrelated (**1**, **2**, and **u**) is statistically significant (p=0.0016).

6.3 Silences

We next looked at how the length of silence between speaker turns (for clean transitions) correlates with boundary tones and relatedness markings. The relatedness markings that we looked at were related-to-last (**0**), and unrelated-to-last (**1 2 u**). Due to the sparseness of data, we clustered silences into two groups, silences less than a half a second in length (labeled *short*), and silences longer than a half a second (*long*). The results are given in Table 4.

Type	Tone			No Tone		
			%			*%*
	Short	**Long**	**Long**	**Short**	**Long**	**Long**
0	160	107	40%	6	11	65%
u,1,2	15	10	40%	8	3	27%

Table 4. Silences

We find that when there is a boundary tone that precedes the new utterance, there is no correlation between relatedness and length of silence (a weighted t-test found the difference in distributions for related-to-last and unrelated-to-last to not be significant, with p=0.851). This suggests that the boundary tone is sufficient as an utterance unit marker and its presence makes the amount of silence unimportant.

In the case where there is no boundary tone, we see that there *is* a correlation between length of silence and relatedness markings. Only 27% of unrelated transitions follow a long pause (the mean silence length was 0.421 seconds, with a standard deviation of 0.411), while 65% of the related transitions follow a long pause (the mean silence length was 1.072 seconds, with a standard deviation of 0.746). Although there are few data points in these two categories, a weighted means t-test found the difference in the distributions to be significant (p=0.014). Thus, long pauses are positively correlated with the previous utterance being grounded. So, if the hearer wishes to reply to speech not ending with a boundary tone, he is more likely to wait longer (perhaps for a completion or repair by the speaker) than otherwise. Thus, silences seem to be a secondary indicator of utterance unit completion, important only in the absence of a boundary tone.

7 Discussion

Our results are still preliminary, due to the small sample of the relevant categories. However, they do show several things convincingly. First, although grounding behavior

is prevalent throughout these problem solving dialogues, there are different degrees to which the speech is grounded. Since adding to the common ground is a prime purpose of conversation, grounding should prove a useful tool for further investigating utterance units and other dialogue phenomena. Second, the claim that utterance units are at least partially defined by the presence of an intonational boundary seems well supported by the conversants' grounding behavior: in addition to serving as a signal for turn-taking, boundary tones also play a role in guiding grounding behavior. Finally, the grounding behavior suggests that pauses play a role mostly in the absence of boundary tones.

There are several ways in which to follow-up this study. First would be just to examine more dialogues and thus get a larger sample of the crucial *related to prior* category, which indicates that the last (partial) unit was deficient but a previous unit was worth grounding. Another possibility would be to look more closely at the overlapped category to gain a handle on projectability, and look closely at the contrast in distributions from the clean transition case. Even our preliminary numbers from Table 3 show that the overlap category contains more cases of no boundary tone, which is what we would expect in a case of disagreement over turn-taking. Finally, it would be very interesting to apply this analysis to other corpora which occur in other domains (which might have different implications with respect to the grounding criterion), other languages (which may differ in how they signal UU completion), and multi-modal communication, in which conversants have other ways to communicate, as well as speech.

7.1 Implications for Spoken Language Systems

In Section 3, we described some of the benefits for spoken language systems that could be derived from having a clear notion of utterance units. The results of this study show that the use of prosodic information will play a key role in the ability to make use of such units. Recognition of prosodic features, especially boundary tones will be crucial to the determination of utterance units. Also, production of appropriate prosodic features will be essential to achieving useful contributions. Since silence can be seen as the termination of an utterance unit in the absence of a boundary tone, it is also important for a production system to avoid pausing where it would not like a user to make such a judgment of unit finality. This could be achieved by using a continuation tune, or perhaps a filled pause.

One of the important features of utterance units is due to the role they play in grounding. They seem to be the basic unit by which human speakers coordinate to ground their contributions. Hence, utterance unit boundaries define appropriate places for the machine to offer backchannel responses, or to check its space of interpretations to determine whether it is appropriate to give positive (e.g., an acknowledgment) or negative (e.g., a repair) feedback.

References

1. James. F. Allen, L. K. Schubert, G. Ferguson, P. Heeman, C. H. Hwang, T. Kato, M. Light, N. Martin, B. Miller, M. Poesio, and D. R. Traum, 'The TRAINS project: a case study in building a conversational planning agent', *Journal of Experimental and Theoretical Artificial Intelligence*, **7**, 7–48, (1995).

2. Leonard Bloomfield, 'A set of postulates for the science of language', *Language*, **2**, 153–164, (1926).

3. Gillian Brown and George Yule, *Discourse Analysis*, Cambridge University Press, 1983.

4. Herbert H. Clark, *Arenas of Language Use*, University of Chicago Press, 1992.

5. Herbert H. Clark and Edward F. Schaefer, 'Contributing to discourse', *Cognitive Science*, **13**, 259–294, (1989). Also appears as Chapter 5 in [4].

6. Starkey Duncan, Jr. and George Niederehe, 'On signalling that it's your turn to speak', *Journal of Experimental Social Psychology*, **10**, 234–47, (1974).

7. Cecelia Ford and Sandra Thompson, 'On projectability in conversation: Grammar, intonation, and semantics'. Presented at the *Second International Cognitive Linguistics Association Conference*, August 1991.

8. Charles Carpenter Fries, *The structure of English; an introduction to the construction of English sentences.*, Harcourt, Brace, 1952.

9. James Paul Gee and Francois Grosjean, 'Saying what you mean in dialogue: A study in conceptual and semantic co-ordination', *Cognitive Psychology*, **15**(3), 411–458, (1983).

10. Francois Grosjean, 'How long is the sentence? Predicting and prosody in the on-line processing of language', *Linguistics*, **21**(3), 501–529, (1983).

11. Derek Gross, James Allen, and David Traum, 'The Trains 91 dialogues', Trains Technical Note 92-1, Department of Computer Science, University of Rochester, (June 1993).

12. M. A. Halliday, 'Notes on transitivity and theme in English: Part 2', *Journal of Linguistics*, **3**, 199–244, (1967).

13. Peter Heeman and James Allen, 'Detecting and correcting speech repairs', in *Proceedings of the 32th Annual Meeting of the Association for Computational Linguistics*, pp. 295–302, Las Cruces, New Mexico, (June 1994).

14. Peter A. Heeman and James F. Allen, 'The Trains spoken dialog corpus', CD-ROM, Linguistics Data Consortium, (April 1995).

15. Peter A. Heeman, Kyung-ho Loken-Kim, and James F. Allen, 'Combining the detection and correction of speech repairs', in *Proceedings of the 4rd International Conference on Spoken Language Processing (ICSLP-96)*, pp. 358–361, Philadephia, (October 1996). Also appears in *International Symposium on Spoken Dialogue*, 1996, pages 133-136.

16. Alon Lavie, Donna Gates, Noah Coccaro, and Lori Levin, 'Input segmentation of spontaneous speech in JANUS: a speech-to-speech translation system', in *Dialogue Processing in Spoken Language Systems*, eds., Elisabeth Maier, Marion Mast, and Susann LuperFoy, Lecture Notes in Artificial Intelligence, Springer-Verlag, Heidelberg, (1997). In this volume.

17. Willem J. M. Levelt, 'Monitoring and self-repair in speech', *Cognition*, **14**, 41–104, (1983).

18. R. J. Lickley and E. G. Bard, 'Processing disfluent speech: Recognizing disfluency before lexical access', in *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP-92)*, pp. 935–938, (October 1992).

19. MADCOW, 'Multi-site data collection for a spoken language corpus', in *Proceedings of the DARPA Workshop on Speech and Natural Language Processing*, pp. 7–14, (1992).

20. M. Mast, R. Kompe, S. Harbeck, A. Kießling, H. Niemann, E. Nöth, E. G. Schukat-Talamazzini, and V. Warnke, 'Dialog act classification with the help of prosody', in *Proceedings of the 4rd International Conference on Spoken Language Processing (ICSLP-96)*, Philadephia, (October 1996).

21. M. Meteer and R. Iyer, 'Modeling conversational speech for speech recognition', in *Proceedings of the Conference on Emphirical Methods in Natural Language Processing*, Philadephia, (May 1996).

22. Shin'ya Nakajima and James F. Allen, 'A study on prosody and discourse structure in cooperative dialogues', *Phonetica*, **50**(3), 197–210, (1993).

23. David Novick and Stephen Sutton, 'An empirical model of acknowledgement for spoken-language systems', in *Proceedings ACL-94*, pp. 96–101, Las Cruces, New Mexico, (June 1994).

24. Bengt Orestrom, *Turn-Taking in English Conversation*, Lund Studies in English: Number 66, CWK Gleerup, 1983.

25. J. B. Pierrehumbert, 'The phonology and phonetics of english intonation', Doctoral dissertation, Massachusetts Institute of Technology, (1980).

26. Massimo Poesio, 'A model of conversation processing based on micro conversational events.', in *Proceedings of the Annual Meeting of the Cognitive Science Society*, (1995).

27. H. Sacks, E. A. Schegloff, and G. Jefferson, 'A simplest systematics for the organization of turn-taking for conversation', *Language*, **50**, 696–735, (1974).

28. Emmanuel A. Schegloff and H. Sacks, 'Opening up closings', *Semiotica*, **7**, 289–327, (1973).

29. Mark Seligman, Junko Hosaka, and Harald Singer, '"pause units" and analysis of spontaneous japanese dialogues: Preliminary studies', in *Dialogue Processing in Spoken Language Systems*, eds., Elisabeth Maier, Marion Mast, and Susann LuperFoy, Lecture Notes in Artificial Intelligence, Springer-Verlag, Heidelberg, (1997). In this volume.

30. K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, 'ToBI: A standard for labelling English prosody', in *Proceedings of the 2nd International Conference on Spoken Language Processing (ICSLP-92)*, pp. 867–870, (1992).

31. J. M. Sinclair and R. M. Coulthard, *Towards an analysis of Discourse: The English used by teachers and pupils.*, Oxford University Press, 1975.

32. Andreas Stolcke and Elizabeth Shriberg, 'Automatic linguistic segmentation of conversational speech', in *Proceedings of the 4rd International Conference on Spoken Language Processing (ICSLP-96)*, (October 1996).

33. Kazuyuki Takagi and Shuichi Itahashi, 'Segmentation of spoken dialogue by interjection, disfluent utterances and pauses', in *Proceedings of the 4rd International Conference on Spoken Language Processing (ICSLP-96)*, pp. 693–697, Philadephia, (October 1996).

34. M. K. Tanenhaus, M. J. Spivey-Knowlton, K.M. Eberhard, and J. C. Sedivy, 'Integration of visual and linguistic information in spoken language comprehension.', *Science*, **268**(3), 1632–34, (1995).

35. David R. Traum and James F. Allen, 'A speech acts approach to grounding in conversation', in *Proceedings 2nd International Conference on Spoken Language Processing (ICSLP-92)*, pp. 137–40, (October 1992).

36. David R. Traum and Elizabeth A. Hinkelman, 'Conversation acts in task-oriented spoken dialogue', *Computational Intelligence*, **8**(3), 575–599, (1992). Special Issue on Non-literal language.