UTTERANCE UNITS AND GROUNDING IN SPOKEN DIALOGUE*

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

Université de Genève and University of Rochester

ABSTRACT

Defining an utterance unit in spoken dialogue 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 signals that have been proposed in the literal, namely prosodic boundary tones and pauses. Preliminary results indicate high correlation between grounding and boundary tones, with a secondary correlation for longer pauses.

1. UTTERANCE UNITS FOR SPOKEN DIALOGUE

An important starting point for the formal study of spoken dialogue is a clear formulation of the basic units of language production and comprehension. For spoken dialogue it has often been claimed that *utterances* rather than *sentences* are the primary object of study [2, 3]. But just what *are* utterances? Following Bloomfield [1], the term *utterance* has often been vaguely defined as "an act of speech." However, 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).

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 [7], for example, uses the term *utterance unit* to denote those chunks of talk that are marked off by a shift of speaker. Some authors (e.g., [11]) also distinguish *speaking-turns* in which new information is conveyed from *backchannel items*, which are short responses such as *ok*, *right*, *yeah*, and *mm-hm*. 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 of the speaker. 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 within the span of a single turn.

From a purely functional point of view, many analysts have found the turn too large a unit for convenient analysis. Sinclair and Coulthard [14], for example, 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, which might be part of different exchanges.

Sacks et. al, [13] present a theory of the organization of turns as composed of *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 may consist of differing types of syntactic contributions, including lexical, phrasal, clausal, and sentential constructions. Much subsequent work on turn-taking (*e.g.*, [5, 11, 6]) 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.

In looking at the relationship of prosody to discourse structure, Nakajima and Allen [10] used four principles to segment turns into utterance units: utterance units correspond to sentences of text, they can correspond to basic speech acts, they are at most a single turn, and they can be marked with a pause of at least 750 msec.

Even though we believe that the basic utterance units can be smaller than individual turns, the turn boundaries are still the easiest place

*Funding for the second author was gratefully received from NSF under Grant IRI-90-13160 and from ONR/DARPA under Grant N00014-92-J-1512. We would also like to thank James Allen.

[1]TECFA, FPSE, Université de Genève, 9 Rte de Drize, 1227 Carouge, Switzerland. Email: David.Traum@tecfa.unige.ch

[2]Computer Science Department, University of Rochester, Rochester, NY 14607, USA. Email: heeman@cs.rochester.edu

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

to recognize utterance unit boundaries. If we assume that the turn is composed of utterance units that are smaller than turns, then the end of a turn should also be an end of an utterance unit. When a new conversant starts to speak, this is partial evidence that he believes that the speech by the previous speaker contained a complete interactional unit. The way he responds can give even more evidence about whether he thinks it is complete. If the new speech builds on or responds to the previous speech, this supports a hypothesis that the previous speech was adequate, whereas if the previous contribution was ignored, this is evidence that it might have been deficient in some manner and not obliging the same type of response.

By examining the relationships of features of the end of the old turn and the beginning of the new turn, we can hope to see which features signal the of end of utterance units. We looked at the *grounding* behavior displayed by the new turn with respect to previous units by the old speaker, contrasting the distribution of types of relationships based on whether boundary tones and pauses were present at the end of the old speaker's turn. In the next section, we discuss grounding behavior and the types of relationships between successive turns. Section 3 includes a description of our corpus and conventions for marking prominent features. Results of this study is presented in 4.

2. GROUNDING AND RELATEDNESS

Clark and Schaefer [4] call the process of adding to common ground between conversants *grounding*. They 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 [15] 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 some acts which present new material, there are *acknowledgment* acts which signal that the current speaker has understood previous material presented by the other speaker, *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.

We use here a much rougher labeling of utterances, since we are not concerned with whether presented material is eventually grounded or not, but merely whether a new turn plays a role in the grounding process, and if so, what previous material it helps to ground. 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 very straightforward to determine whether or not it performs some 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 which follow backchannels and other turns which consist solely of these signals, there is no real content for the new turn to be related *to*. We also allow categories 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. Table 1 summarizes this coding scheme, as used to mark turn-initial relatedness to utterance units from the previous turn. Examples are also presented in the next section.

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 utterance one **previous** to the most recent but *not* related 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 utterance unit (which would be marked **0**), or to an utterance unit 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

3. DATA

To study the grounding and turn-taking phenomena, we analyzed and labeled a corpus of problem-solving spoken dialogs in which the conversants had no visual contact. Since this corpus contains dialogues in which the conversants work together in solving the task, the grounding criterion was fairly high, and a high degree of grounding behavior is expected. For our current study, we looked at 26 separate dialogues[4] ranging in length from 50 to 500 seconds. This corpus totaled over 6000 seconds of spoken dialogue, comprising 1366 turn transitions.

3.1. Prosodic Markings

We mark two kinds of prosodic information which has been used as indication of utterance unit boundaries. First, we consider Pierrehumbert's intonation phrase (IP) [12]. 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

[4] Selected at random from the TRAINS-93 Dialogues [8, 9].

there was no completed IP in the last turn by the previous speaker then the entire turn up to the transition point is counted as a partial IP. We also note the amount of silence between the end of one turn and the resumption of the next to see if pause length is a significant indicator of utterance unit boundaries. This is marked in the examples below by indicating the time in seconds inside square brackets.

3.2. Relatedness Markings

Each turn-transition was marked, using the scheme described in Table 1, as to how the initial installment of the new turn related to the previous completed or uncompleted IPs produced by the other speaker. For cases of overlapping speech, the current ongoing installment by the continuing speaker is considered the most recent (partial) IP. 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.

3.3. Examples

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.[5]

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
```

The following example shows some of the other categories.[6] The second turn by U shows the **u-e** category, since the previous turn was just "okay". This turn also ends without a boundary tone. S's second turn seems to be some sort of clarification attempt, but it is not clear if it is at all related to the content of U's previous utterance. U's final utterance is merely a continuation of his own previous utterance and is unrelated to the last installment by S (which also has no final boundary tone).

Example: d93-16.2 utt27-31

```
              % [.73]
U: then do that
   e  % [.38]
S: okay
      u-e                                          [1.54]
U: um what is tank- what is engine okay engine two is
   ?                 [.26]
S: you have you told
   u                       %
U: it's picking up the oranges
```

[5] File [IMAGE A731G01.GIF] shows wave form, pitch contour and annotations for this excerpt, while the sound is in [SOUND A731S01.WAV].

[6] In files [IMAGE A731G02.GIF] and [SOUND A731S02.WAV].

4. RESULTS

4.1. Prevalence of Grounding Behavior

Tabulating the markings on the beginning of each stretch of single speaker speech 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 (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, and only 3% are clearly unrelated and do not follow an acknowledgment.

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

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 in [15] 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.

4.2. 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 [13]: 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.

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.

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

4.3. 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, *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 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, in the case in which there is no preceding boundary tone, long pauses are positively correlated with relatedness to the previous utterance, and thus long silences seem to be a secondary indicator of utterance unit completion, important only in the case of a lacking boundary tone.

5. CONCLUSIONS

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 dictating grounding behavior. Finally, the grounding behavior suggests that pauses play a role mostly in the absence of boundary tones.

6. REFERENCES

1. Leonard Bloomfield. A set of postulates for the science of language. *Language*, 2:153–164, 1926.
2. Gillian Brown and George Yule. *Discourse Analysis*. Cambridge University Press, 1983.
3. Herbert H. Clark. *Arenas of Language Use*. University of Chicago Press, 1992.
4. Herbert H. Clark and Edward F. Schaefer. Contributing to discourse. *Cognitive Science*, 13:259–294, 1989. Also appears as Chapter 5 in [3].
5. Starkey Duncan, Jr. and George Niederehe. On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, 10:234–47, 1974.
6. Cecilia Ford and Sandra Thompson. On projectability in conversation: Grammar, intonation, and semantics. presented at the Second International Cognitive Linguistics Association Conference, August, 1991.
7. Charles Carpenter Fries. *The structure of English; an introduction to the construction of English sentences.* Harcourt, Brace, 1952.
8. Peter A. Heeman and James Allen. The TRAINS 93 dialogues. TRAINS Technical Note 94-2, Department of Computer Science, University of Rochester, 1994.
9. Peter A. Heeman and James F. Allen. The Trains spoken dialog corpus. CD-ROM, Linguistics Data Consortium, April 1995.
10. Shin'ya Nakajima and James F. Allen. Prosody as a cue for discourse structure. In *Proceedings 2nd International Conference on Spoken Language Processing (ICSLP-92)*, pages 425–428, October 1992.
11. Bengt Orestrom. *Turn-Taking in English Conversation*. Lund Studies in English: Number 66. CWK Gleerup, 1983.
12. J. B. Pierrehumbert. The phonology and phonetics of english intonation. Doctoral dissertation, Massachusetts Institute of Technology, 1980.
13. H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50:696–735, 1974.
14. J. M. Sinclair and R. M. Coulthard. *Towards an analysis of Discourse: The English used by teachers and pupils.* Oxford University Press, 1975.
15. 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)*, pages 137–40, October 1992.