Intonation as a Presentational Resource in Conversation

Susan E. Brennan

Psychology Department
State University of New York
Stony Brook, NY 11794
brennan@psych.stanford.edu

Workshop on Prosody in Natural Speech, U. Penn, August 5-12, 1992

KEYWORDS: *Intonation, prosody, mutual knowledge, communication, conversation, collaboration, grounding, backchannels*

1. Communication as collaboration

The collaborative view of language use holds that speakers and addressees are jointly responsible for contributions to conversations (Clark & Wilkes-Gibbs, 1986). They do not simply produce and comprehend utterances autonomously; instead, they achieve a state of mutual knowledge by exchanging evidence that they've understood one another (Brennan, 1990; 1992; Clark & Wilkes-Gibbs, 1986; Schober & Clark, 1989). This process of exchanging evidence is called *grounding*. Evidence of understanding can take many forms, including an appropriate second part in an adjacency pair, such as the answer to a question (Levinson, 1983), a relevant next turn in a conversation (Clark & Schaefer, 1987, 1989), backchannels (Yngve, 1970), or eye contact (Goodwin, 1981). The evidence used to ground utterance meaning can be provided verbally or visually (Brennan, 1990; 1992), and conversation is shaped in part by by the resources available for grounding in a particular communication medium (Clark & Brennan, 1991; Whittaker, Brennan, & Clark, 1991). And conversational partners appear to set higher or lower *grounding criteria* for the amount of evidence they seek and provide before concluding they understand one another well enough for the purpose at hand (Clark & Wilkes-Gibbs, 1986; Wilkes-Gibbs, 1986).

In the *contribution model* of Clark and Schaefer (1987, 1989), every contribution to a conversation has two phases: a presentation phase and an acceptance phase. Every utterance is itself a presentation; it does not become a contribution to the conversation until its acceptance phase is complete, that is, until the addressee provides evidence that he believes he understands what the speaker meant, and the original speaker ratifies that evidence. So even though a speaker may have an intention in mind when she presents an utterance, her utterance does not stand as a contribution to the conversation until she has evidence from her addressee that he has understood[1]. For instance, consider this example from the Lund corpus (Svartvik & Quirk, 1980), where A and B are two people engaged in conversation:

```
A: is term OK

B: what

A: is term all *right*

B: *yes* it seems all right so far
   . touch wood
```

Here, A presents an utterance that may have been intended to function as a question. But after A's first turn, B's utterance provides evidence that he does not yet understand. There are of course many possibilities, including these – perhaps he didn't hear her, perhaps he didn't understand what she meant by "OK," perhaps he didn't catch the word "term." A repairs her original presentation by repeating it with a word change - from "OK" to "all right." The end of A's utterance is overlapped by the beginning of B's (overlapping parts are indicated by the asterisks). Since B begins his utterance early, it could be that the problem was with the word "term." In any event, it is not until the fourth turn, when A hears B's relevant response, that she can surmise that her question has been properly understood by B. At this point A can go on with another relevant presentation. In doing so, she communicates to B that she is satisfied that the acceptance phase for the utterance she first presented is complete; after that, a contribution to the conversation has been made.[2]

[1] For clarity, I will use the convention that speakers are female and addressees are male.

[2] See Brennan & Cahn, 1992, for a discussion of the temporal asymmetry in A's and B's roles in producing a contribution to the conversation.

At each moment in a conversation, an individual can provide evidence to a partner, or seek evidence from that partner (Brennan, 1990; 1992). The exchange of evidence happens in a systematic way: A speaker presents constituents of an appropriate size, depending on the grounding criterion and on the communication medium, and an addressee typically provides appropriate evidence of understanding as soon as he concludes that he has understood a speaker's presentation well enough for current purposes. If an addressee believes that he does not understand, he will withhold the expected positive evidence, or provide explicit negative evidence (e.g. with a clarification question, a puzzled frown, etc.). If the expected evidence of understanding from an addressee is not forthcoming, a speaker will pursue it (Brennan, 1990; 1992; Pomerantz, 1984).

In this paper I present some data about the intonational resources that people can use for grounding the meanings of utterances. My claim is that intonation *not only* conveys information about syntactic constituents (Cruttendon, 1986) and the speaker's intention (Sag & Liberman, 1974; Liberman & Sag, 1974; Pierrehumbert & Hirshberg, 1990), *but also* can be used to manage the exchange of evidence between two people in conversation, en route to achieving mutual understanding. In particular, I examine phrase final rising intonation. It has been proposed by some that such intonation serves an interactional purpose (Brennan, 1990; McLemore, 1991), e.g. to elicit the attention of addressees, or to pursue a response. I will bring behavioral evidence to bear on the hypothesis that speakers use rising intonation to actively seek evidence of understanding from their addressees.

2. Intonation as a presentational resource

My corpus consists of stereo recordings of conversations about map locations. Pairs of people did a matching task using pictures of the same map displayed on two computers networked together. Since the task was to get both of their cursors located in the same target location on the map, the degree to which they understood one another was indexed by the distance between their cursors. Throughout their conversation, a log was kept of cursor position, and this log was later synchronized with the conversational transcript. This technique enabled a continuous online measure of understanding in conversation.

2.1. Method

Subjects were 24 Stanford graduate students between the ages of 21 and 32, all native speakers of American English. They participated as same-sex pairs who had never met one another before. There were eight women and 16 men, from 13 different academic departments. They were recruited through posted advertisements or electronic bulletin boards, and participated in exchange for a small honorarium.

Pairs of subjects in adjoining cubicles used computers networked together to do a matching task. Each partner was seated in front of an identical computer graphics display of a map. Each display had a small icon of a car. The task was for one person (the director, D) to convey a target location to the other (the matcher, M), and for the matcher to position a car icon in the target location by dragging it with his mouse. The task was done 80 times by each pair. They could talk to each other as much as they liked, but they could not see each other.

There were two experimental conditions, *visual evidence* and *verbal-only evidence*. That is, in half of the trials, the director could see the position of the matcher's icon on the screen, and so had visual evidence of exactly what the matcher understood; in the other half of the trials, there was no visual evidence (the director could see only her own icon). After the matcher "parked" his car in a location by clicking his mouse, the director initiated a new trial by clicking on her icon, which then moved by itself to a new preprogrammed location. Subjects' displays were always identical, except for their icon positionss. Maps of the Stanford University campus and of Cape Cod were used as graphic backgrounds for the trials. After every block of ten trials, the pair of subjects alternated evidence conditions, maps, or director/matcher roles.

2.2. Analysis

Speech transcripts. The conversations of six of the 12 pairs of subjects were chosen randomly for detailed transcription, yielding 480 descriptions of map locations. These descriptions were transcribed in segments that corresponded roughly to one phonemic clause per line (that is, a short sequence of words separated by a pause, and generally containing one primary pitch accent (Rosenfeld, 1987; see also Boomer, 1978; Dittmann & Llewellyn, 1967)). Each line was punctuated in order to categorize its clause-final prosody approximately: "." for final pitch lowering, "?" for final rising, "," for the end of a tone unit (if mid-clause) or else for list-like intonation (when at the end of a clause), "-" for a sudden self-cutoff on a level pitch, and no punctuation for clauses fitting none of these criteria. The clauses sometimes had extreme final lengthening, or drawled syllables, which were denoted by ":" following the letter that

most closely matched the sound being drawn out (ye:s for "yeeees," vs. yes: for "yesss"). Overlapping speech was transcribed using single or double asterisks to enclose the beginning and ending of both stretches of simultaneous talk. Unintelligible speech was enclosed in brackets. All transcripts were double checked for accuracy.

In order to conduct a detailed analysis of individual conversational interchanges, I took a random sample of 48 of the 480 transcribed interchanges. One item (that is, one location on the map) was chosen at random from each cell in the counterbalanced design of the experiment. In this smaller sample, each pair of subjects contributed interchanges concerning the same 8 map locations. I then coded whether or not the director in each interchange used the final rising intonation typical of question intonation in presenting a description, in the intial period before the matcher had made any verbal response.

Action transcripts. During each trial, the x and y coordinates of matcher's icon were recorded and time-stamped, to provide a record of the matcher's understanding of the location of the target. For the small sample of 48 trials, the distance between the matchers' icon and the target location (the director's icon) was plotted over time, to provide a visible display of on-line understanding (convergence between the two icons) in the conversational interchange.

Then the 48 action transcripts were synchronized with the speech transcripts. A naive coder, with copies of the language transcripts in front of her, listened to the tapes of the conversational interchanges using a videotape player that was equipped with a seconds counter. For each of the 48 trials in the sample, she zeroed the counter at the start of the trial (marked by a short beep) and recorded an integer at every 1.0 second interval by writing the integer over its corresponding word on the transcript. It took several passes over the tapes to record the seconds intervals and to check the synchronization of each trial. We estimate that this procedure was accurate well within a half-second. Synchronization of these action transcripts with the speech transcripts is shown using matching superscripts over the speech transcripts and the time-distance plots (see Figure 1).

Did the directors use rising intonation differently in their presentations to matchers when they could see what the matchers were doing vs. when they could not? The collaborative view predicts that they should; the exchange of evidence via intonation should be managed differently in a medium where visual evidence is available than in one where visual evidence is not (Clark & Brennan, 1991). I coded whether or not D used final rising intonation, often associated with questions in English, in any of the descriptions she uttered initially in an interchange, *before* the point where she got a verbal response from M. Then I coded whether or not M had moved his icon before responding verbally or before D's use of question intonation (whichever came first). The expectation was that D could use either M's verbal response, or M's icon movement (in the visual evidence condition) to conclude that M had understood her description of a map location.

2.3. Results and discussion

While the baseline frequency of using final rising intonation was the same across both evidence conditions – that is, directors were just as likely to use final rising intonation before the matcher's first verbal response in the verbal-only condition as they were in the visual condition (58.3 % to 41.7 %, n.s.), final rising intonation was distributed differently in the two evidence conditions. D's use of final rising intonation was related to whether M had moved his icon yet in the visual evidence condition, but not in the verbal-only evidence condition. With visual evidence, there was a correlation between D's use of final rising intonation and M's lack of icon movement was ($r_\phi = .48$, $p < .02$). That is, when the directors could monitor the matchers' icon visually and the matchers hadn't yet made any progress toward the target during the directors' initial descriptions, the directors were likely to use final rising intonation, possibly to pursue a response from the matchers. There was no such systematic relationship in the verbal evidence condition, where the directors weren't able to see whether the matchers had moved yet ($r_\phi = .17$, $p < .50$).

Let us take a closer look at the relationship between D's intonation and M's icon movement. Consider this example where D did *not* use any final rising intonation.

```
Example 1:

D: ok
   now we went
   south,
   we're
   about halfway down the screen
   in the electronics lab,
   we're in the southern most
   wing of the electronics lab
   good
   good

M: I think that's where my office is.
(Pair 3, Location 33, Visual evidence condition)
```

In this example, M began to move his icon immediately after D said "south" and made fairly steady monotonic progress toward the target (except for during the point when D was pronouncing "southern most" – see the time-distance plots for this example and others in the Appendix). M arrived at the target just before D 's second use of the word "lab," and D, since she had visual evidence that M's hypothesis about what she meant was correct, acknowledged this with "good good." A similar pattern was found in the next example (Example #2, Appendix).

Example 2:

D: uhh
Terman Engineering
just to the lower right of the five.

M: ok, Terman?
to the lower right?

D: yah
bingo.

M: right here?

D: right there.
(Pair 4, Location 25, Visual evidence condition)

In this example, M also made early progress toward the target, and D could monitor this and did not use final rising intonation. M sought further evidence about D's meaning by taking a verbal turn. In addition, this pair explicitly acknowledged having visual evidence by using a deictic strategy with: "bingo. - right here? - right there."

In the next example, M did not start moving his icon until after D said "Terman Engineering?"

Example 3:

D: um
now it's over:
near the right edge
near Terman Engineering?
right next to the number five?
below and to the right
of of the five?
right there.
(Pair 2, Location 25, Visual evidence condition)

As M started moving, D used final rising again. At the point where D said "number five?" M had gone just past the correct location. D provides a more detailed description: "below and to the right of of the five?" and about half a second after that, M's icon arrives in the target location.

Directors also used final-rising intonation when they could *not* see the matchers' icon. In the next example, from the verbal-only evidence condition, D happened to use final rising intonation before M moved. Here, D had started out with a very general description, "you're going down," and then explicitly pursued information from M before providing a more explicit description:

Example 4:

D: ok
you're going down to
uhh
do you know where
the electronic
labs are?
S E L?

M: yah I see it

D: ok....[continues]
(Pair 2, Location 33, Verbal evidence condition)

In this example, M responded both verbally and by starting to move his icon at roughly the same time, just after D's first final rising intonation, and as D finished saying "S E L?" Despite examples like this one, there was no significant correlation between D's intonation and M's icon movement in the verbal evidence condition, nor would we expect there to be, since D had no direct evidence of whether M had moved yet.

3. Conclusions

Conversations can include both verbal and visual evidence, as two people try to reach a point where they believe they understand one another well enough for their current purposes. The grounding process is both flexible and opportunistic; when visual evidence is available, it can "fill in" for verbal evidence – that is, it can substitute for a verbal turn in the conversation (Brennan, 1990; 1992). When a speaker gets an appropriate response from an addressee, she can conclude that the utterance she presented has been accepted; when she doesn't get the evidence she expects, she can actively seek such evidence.

In this study, speakers often used final rising intonation when their addressees had not yet provided any evidence

of understanding. It is possible that speakers did this in order to pursue a response from their addressees, since it was correlated with the addressee's lack of movement in the visual condition, and uncorrelated with the addressee's lack of movement in the verbal condition. During the initial presentation of map location descriptions, the baserates for final rising intonation were just as high in the verbal condition as in the visual one, but in the verbal condition it was distributed randomly, (or at least it was not predicted by the addressee's icon movement).

There is a difference between when an addressee has a hypothesis about what a speaker means, and when they can conclude that they understand one another (Schober & Clark, 1989). In this study, the time-distance plots of the convergence of the two icons show a distinct elbow when the matcher's icon arrives within close range of the director's icon. When the director can see the matcher's icon, the conversational interchange is brought to a rapid conclusion, often by the director using deictic means (e.g. "park right there!"). When the director cannot see the matcher's icon, there is a relatively long period where they must continue grounding, until they can conclude they've reached a state of mutual understanding (Brennan, 1990; 1992).

For the future, I plan to examine additional prosodic aspects of these dialogues having to do with turn placement. These include early or overlapping turns that may provide early evidence that M believes he has understood D's presentation, and also pauses that may mark a presentation as problematic and in need of repair (see Levinson, 1983; Jefferson, 1989).

4. References

Boomer, D. S. (1965). Hesitation and grammatical encoding. *Language and Speech*, 8, 148-158.

Boomer, D. S. (1978). The phonemic clause: Speech unit in human communication. In A. W. Siegman and S. Feldstein (Eds), *Nonverbal behavior and communication.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Brennan, S. E. (1990). Seeking and providing evidence for mutual understanding. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Brennan, S. E. (1992). On the time course of understanding in conversation. Manuscript in submission.

Brennan, S. E. and Cahn, J. (1992). An architecture for contributions and repair in a natural language interface. Manuscript in preparation.

Clark, H. H. and Brennan, S. E. (1991). Grounding in communication. In L.B. Resnick, J. Levine, and S.D. Teasley (Eds.), *Perspectives on Socially Shared Cognition.* Washington, DC:APA.

Clark, H. H. and Schaefer, E. F. (1987). Collaborating on contributions to conversations. *Language and Cognitive Processes*, 2, 19-41.

Clark, H. H. and Schaefer, E. F. (1989). Contributing to discourse. *Cognitive Science*, 13, 259-294.

Clark, H. H. and Wilkes-Gibbs, D. (1986). Referring as a collaborative process. *Cognition*, 22, 1-39.

Cruttenden, A. (1986). *Intonation.* Cambridge, UK: Cambridge University Press.

Dittman, A. T. and Llewellyn, L. G. (1967). The phonemic clause as a unit of speech decoding. *Journal of Personality and Social Psychology*, 6, 341-349.

Goodwin C. (1981). *Conversational organization: Interaction between speakers and hearers.* New York: Academic Press.

Jefferson, G. (1989). Preliminary notes on a possible metric which provides for a 'standard maximum' silence of approximately one second in conversation. In D. Roger and P. Bull (Eds.), *Conversation.* Philadelphia: Multilingual Matters Ltd.

Levinson, S. (1983). *Pragmatics.* Cambridge: Cambridge University Press.

Liberman, M. and Sag, I. A. (1974). Prosodic form and discourse function. In *Papers from the Tenth Redional Meeting*, Chicago Linguistics Society, Chicago, IL.

Maclay, H. and Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, 15, 19-44.

McLemore, C. (1991). The pragmatic interpretation of English intonation: Sorority speech. Unpublished doctoral dissertation, University of Texas, Austin, TX.

Pomerantz, A. (1984). Pursuing a response. In J. M. Atkinson and J. Heritage (Eds.), *Structures of social action.* Cambridge: Cambridge University, University Press.

Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In: P. R. Cohen, J. Morgan, and M. Pollack (Eds.). *Intentions in Communication.* Cambridge, MA: MIT Press.

Rosenfeld, H. M. (1987). Conversational control functions of nonverbal behavior. In A. W. Siegman and

S. Feldstein (Eds.), *Nonverbal behavior and communication.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Sag, I. A. and Liberman, M. (1975). The intonational disambiguation of indirect speech acts. In *Papers from the Eleventh Redional Meeting*, Chicago Linguistics Society, Chicago, IL.

Svartvik, J. and Quirk, R. (Eds.)(1980). *A corpus of English conversation.* Lund, Sweden: Gleerup.

Whittaker, S. J., Brennan, S. E., and Clark, H. H. (1991). Coordinating activity: An analysis of interaction in computer-supported cooperative work. Proceedings, CHI '91, Human Factors in Computing Systems, ACM Press, pp. 361-367.

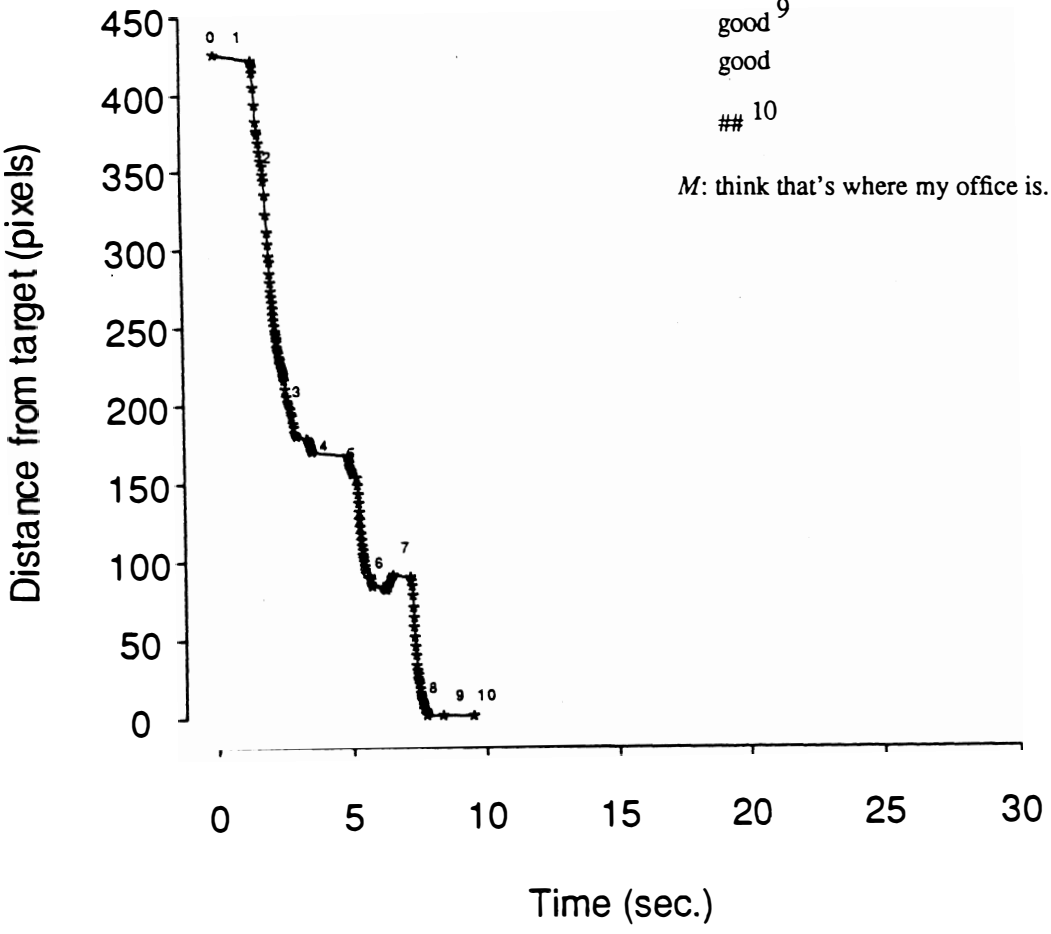
Wilkes-Gibbs, D. (1986). Collaborative processes of language use in conversation. Unpublished doctoral dissertation, Stanford University, Stanford, CA.

Yngve, V. H. (1970). On getting a word in edgewise. In *Papers from the Sixth Regional Meeting of Chicago Linguistic Society* (pp. 567-578). Chicago: Chicago Linguistic Institute.

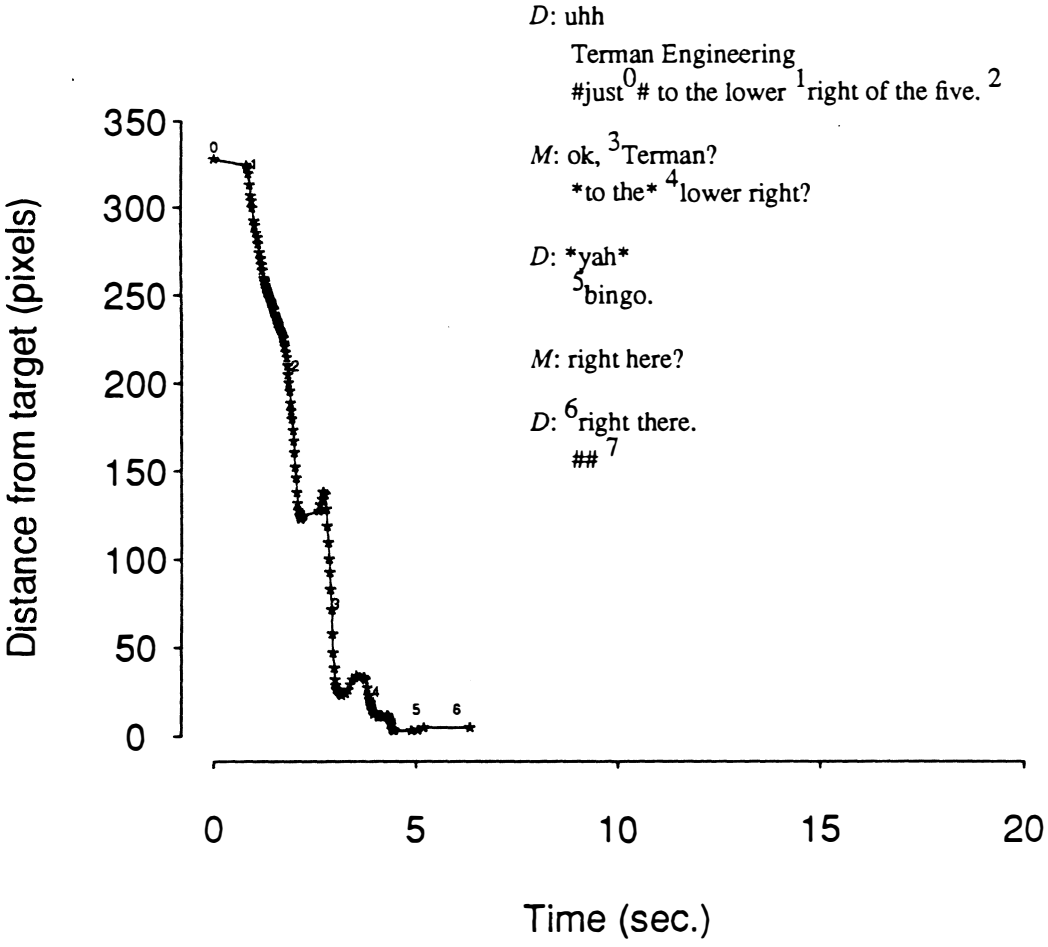
Appendix

D: ok
#[0]now we# went
[1]south,
[2]we're
about halfway down the [3]screen,
in the elec[4]tronics lab, [5]
we're in the [6]southern most
[7]wing of the electronics [8]lab,
good [9]
good
[10]

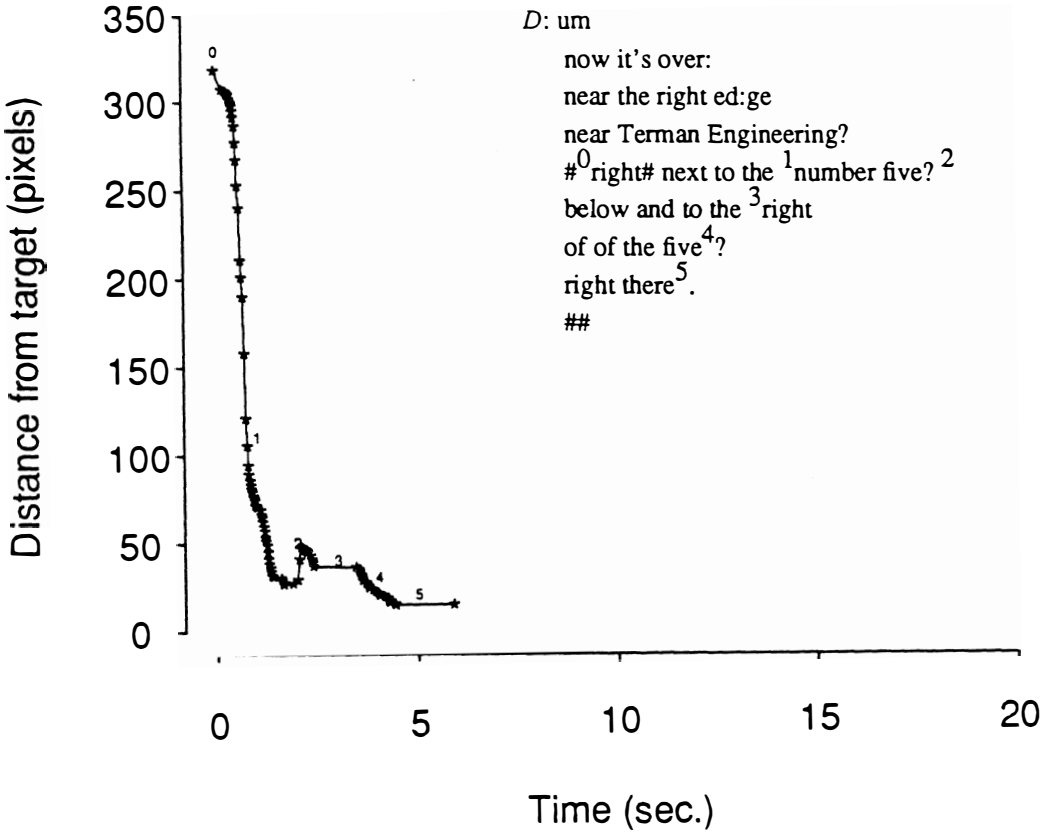
M: think that's where my office is.



Example 1 Visual evidence, Stanford map, Item 33, Pair 3



Example 2 Visual evidence, Stanford map, Item 25, Pair 4



Example 3 Visual evidence, Stanford map, Item 25, Pair 2

D: ok
you're going #[0]down to:#
[1]uhh
do you know wher:e
the [2]electronic
[3]labs are?
[4]S E L?*

M: *yah I see [5]it

D: ok,
[6]it's:
[7]right above
the:e
[8]TRON
[9]in
Elec[10]tronics,
and [11]it's
on that lower E[12]
part of the building. [13]

M: wait[17], it's ju- just above[15] thuuh

D: [16]Tron.

M: the [17]tronn?

D: yah, [18]on the building?

M: [19]what is the tron? [20]

D: oh-
in the
[21]elecTRONics labs?

M: [22]oh,
oh the TRON. [23]

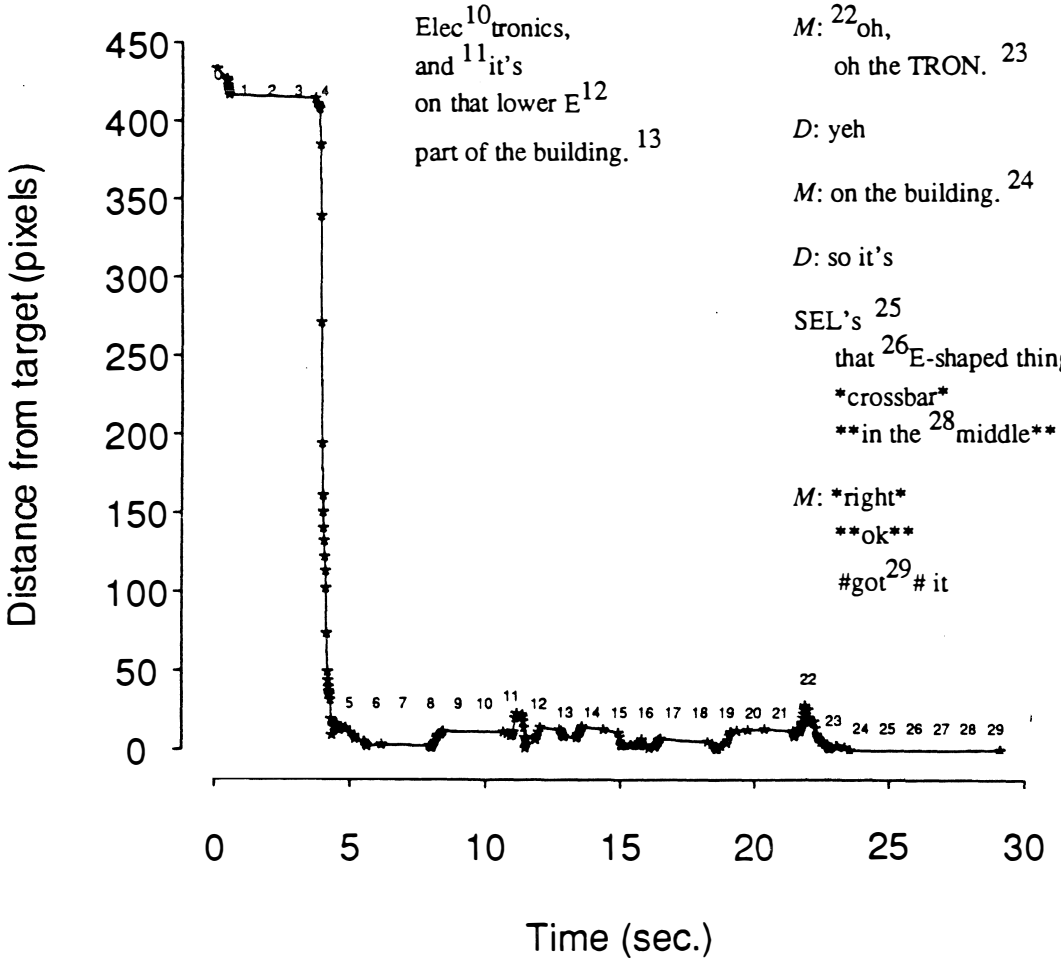
D: yeh

M: on the building. [24]

D: so it's

SEL's [25]
that [26]E-shaped thing's [27]lower
crossbar
in the [28]middle

M: *right*
ok
#got[29]# it



Example 4 Spoken evidence, Stanford map, Item 33, Pair 2