Display Acts in Grounding Negotiations

Yasuhiro Katagiri* **and Atsushi Shimojima†**

*ATR Media Integration & Communications Research Laboratories
2–2–2 Hikari Seika Soraku Kyoto, 619-0288 Japan
katagiri@mic.atr.co.jp

†Japan Advanced Institute of Science and Technology
1–1 Asahi Tatsunokuchi Nomi Ishikawa, 923-1292 Japan
ashimoji@jaist.ac.jp

Abstract

Display utterances are those utterances, such as verbatim rephrases or cooperative completions, that exhibit the speakers' construals of their partners' previous utterances. Drawing on both empirical analyses and theoretical considerations, we study the ways they perform the acts of acknowledgment and repair-request. We show that (1) these grounding acts, when performed by display utterances, are *not* autonomous acts whose choices are under the control of the speakers, yet (2) they are instantaneous acts whose types are fixed at the time of utterance. Accordingly, the grounding model proposed in this paper reconciles the non-autonomous nature of display utterances (Clark, 1996) with the necessity for on-line classifications of the grounding acts (Traum, 1994).

1. Introduction

What one says may not be immediately shared in a natural conversation due to various possibilities including communication error and disagreement. The process of grounding (Clark and Schaefer, 1989; Traum, 1994) is a dialogue process in which a piece of information contributed by one speaker becomes established in the common ground between dialogue participants. Each utterance made in a dialogue is considered to perform a particular *grounding act*, depending on the contribution it makes to the overall process. For example, Clark and Schaefer (1989) characterize the fundamental structure of a grounding process as a sequence of two grounding acts, called *presentation* and *acceptance*; Traum (1994) proposes a more fine-grained classification of grounding acts, including *initiation*, *continuation*, *acknowledgment*, and *cancel*.

Given this view, it is but a short step to make the following assumption on the nature of grounding acts:

Autonomy: The type of grounding act performed by an utterance is under the control of the speaker.

In this paper, we look at the class of utterances that have been called "displays" (Clark, 1996), and show that the proper treatment of their grounding functions calls for a revision of this autonomy assumption. An utterance is a *display utterance* if it somehow exhibits a part of what the speaker has construed out of a preceding utterance by a different speaker. Therefore, the grounding act attributed to a display utterance is typically an acknowledgment and sometimes a repair-request. Drawing on an analysis of our data on language-prosody interaction in "echoic responses" in real dialogues, we show that the kind of grounding act being performed by a display utterance depends on contextual factors beyond the speaker's access or control, and hence the acts of acknowledgment and repair-request, when performed with display utterances, are not generally autonomous acts.

We then propose a new way of looking at the grounding functions of display utterances, which replaces the autonomy assumption with the following slightly different assumption:

Instantaneity: The type of grounding act performed by an utterance is fixed at the time of the utterance.

Clark and Schaefer (1989) and Clark (1996) emphasize the non-autonomy or jointness of grounding processes; one of the main appeals of the grounding model in Traum (1994) is that it retains the instantaneity of grounding acts in the above sense. The view to be proposed in this paper demonstrates that autonomy and instantaneity are different matters, and giving up one does not entail giving up the other—in other words, we can pursue Traum's program while doing justice to the jointness of certain grounding acts.

2. Correctness of Construals

A quintessential example of a display utterance is an *echoic response*. In the following example, B's response directly displays "on Chestnut Street" as a part of what B has extracted out of A's preceding utterance.

(1) A: Sue's house is on Chestnut Street.
B: On Chestnut Street.

A *collaborative completion* can also be considered as a display:

(2) A: The next meeting will be Tuesday next week
B: at the same time in the same room, right?
A: Yeah.

Unlike the case of an echoic response, the information displayed in B's utterance has never been explicitly stated by A; it is instead something B has inferred from A's somewhat incomplete utterance, as a part of what A had intended to convey.

A display of one's construal may be done rather indirectly, as in the following example of a third-turn repair:

(3) A: Did Mary go to the party?

B: I didn't see her last night.
A: No, I meant John's party on last Thursday.

Here, B's utterance displays her construal that A's preceding question refers to the party last night, and this is why A can ever correct B's construal of the intent of her preceding question.

Our notion of display encompasses what Clark and Schaefer (1989) call "display" and "demonstration"; Traum (1994) discusses display utterances mainly as a way to perform an acknowledgment; furthermore, our notion is almost coextensional with "public display" envisioned by Clark (1996). Thus, display acts have often been at the center of attention in discussions on grounding phenomena. Yet, the exact ways they perform their grounding acts have hardly been incorporated into a systematic theory of grounding.

In fact, a problem occurs if we try to understand the grounding functions of display utterances with a straightforward application of the grounding model of Clark and Schaefer (1989). According to Clark and Schaefer, a speaker B is said to *accept* an utterance u by a speaker A under the following conditions:

1. B gives evidence e' that he believes he understands what A means by u.
2. B does so on the assumption that, once A registers evidence e', he will also believe that B understands.

Clark and Schaefer cite display utterances as one of the major ways of performing an acceptance, but if we strictly apply the above conditions, few instances of display can actually perform an acceptance. Take (1) for example. Here, B may believe that he understands what A means by his utterance, and the timing and prosody of B's echoic response may indeed signal this belief on B's part. Therefore, B's utterance may satisfy condition 1 above.

Yet, it is unlikely that condition 2 holds at the time of this utterance. To B, there is always a possibility that his echoic response may be an incorrect repeat of A's original utterance. If that is the case, B's utterance will certainly fail to lead A to believe that B understands; it may even lead A to believe that B *doesn't* understand. Accordingly, unless B is sure that he has correctly repeated A's utterance, B cannot assume that her utterance will lead A to believe that B understands. As Clark (1996) shows, however, a display utterance is essentially a part of a joint-construal activity, where a speaker displays her construal of a previous utterance to allow another speaker to check its correctness. Clearly, B would not engage in such an activity if B were sure that his construal of A's utterance is correct. Condition 2 is therefore hardly satisfied by a display utterance.

The crucial point is that a display utterance always comes with the risk of exhibiting an incorrect construal of a previous utterance, while the correctness of the displayed construal is usually beyond the control of the speaker. (Condition 2 goes against this nature of display utterances by effectively requiring the speaker to know the correctness of her construal.) Now, generally, in order for a display utterance to function as an acknowledgment, the displayed construal must be correct; furthermore, an utterance that exhibits an incorrect construal almost always prompts a repair from the original speaker. This shows that, when display utterances are involved, the act of acknowledgment and the act of repair-request are *not* autonomous acts: the speaker has no absolute control over which act she is performing with her display utterance, inasmuch as she cannot perfectly predict whether she is demonstrating understanding or misunderstanding.

Note that such a radical uncertainty does not exist if the speaker uses a more explicit form of acknowledgment or repair-request. Consider the following modifications of example (1):

(4) A: Sue's house is on Chestnut Street.
B: Uh huh. (or B: What?)

Unlike a display utterance, the "uh huh" in (4) does not show the content of B's construal of A's utterance; B withholds that information, and accordingly gives no concrete clue for A to evaluate the correctness of B's construal. Therefore, A cannot help but rely on the convention that "uh huh" is uttered only when the speaker understands (or believes to understand) the preceding utterance. For this reason, B can assume that his utterance of "uh huh" leads A to believe that B correctly understands, satisfying condition 2 in Clark and Schaefer's model (1989). Also, precisely because the content of B's construal is withheld, there is no chance that it is revealed to be incorrect. Accordingly, barring exceptional circumstances, B's utterance of "uh huh" never functions as a repair-request. For a parallel reason, B can be confident that that her utterance serves as a repair-request rather than an acknowledgment when she utters "What?" In both instances, B has control over what type of grounding act she is performing with her own utterance.

3. Integration of Construals

Clearly, whether the construal exhibited in a display utterance is correct is a dominant factor that affects the type of grounding act being performed. But it is certainly not the only factor: the degree in which the speaker is convinced of her construal, often signaled by the timing and prosody of her speech, also seems to be a strong factor.

Consider the echoic response in (1) again. Intuitively, if it were made in a falling tone without any delay, A would feel that B has integrated the displayed construal ("on Chestnut Street") well in her body of knowledge; A would be sure of the success of grounding and perhaps go on to the presentation of the next item of information. On the contrary, if B's response were made in a rising tone with a considerable delay, A would doubt that B has adequately integrated the information; A would be prompted to restate or rephrase the information to supplement the grounding failure. This suggests that, even when confined to the case where the displayed construal is correct, an echoic response shifts its grounding function between an acknowledgment and a repair-request, *depending on* the speaker's integration signaled by the timing and prosody of the utterance.

In fact, our previous studies on the functions of echoic responses (Shimojima et al., 1998; Shimojima et al., 1999) lend empirical supports to this intuition. We conducted a

corpus-based observational study and three experiments in the following procedures:

Observational study Instances of echoic responses were extracted from three samples of task-oriented spoken dialogue data, and the correlation between their temporal and prosodic features and the raters' assessments of the speakers' integration were examined.

Experiment 1 Acoustically manipulated speech samples of echoic responses were presented to subjects who were asked to evaluate the speakers' integration.

Experiment 2 Instances of echoic responses extracted from our corpus were presented to subjects who were asked to judge the grounding functions being performed. They were to choose from "acknowledgment" and "request repair."

Experiment 3 The same stimuli were used, while the subjects who were asked to judge the most appropriate response to each instance.

The observational study and Experiment 1 provided an evidence that the prosodic and temporal features of an echoic response indeed signal the degree of the speaker's integration. Specifically, a long delay, a rising boundary tone, a high pitch, or a low tempo indicates a high integration, a short delay, a falling boundary tone, a low pitch, or a high tempo indicates a low integration.

Furthermore, Experiments 2 and 3 showed the correlation between the integration rates associated with echoic responses and the subjects' judgments on their grounding functions: when the integration is high, the subjects tend to take it as an acknowledgment and to choose a response appropriate to an acknowledgment, whereas in the case of a low-integration echoic response, the subjects tend to take it as a repair-request and to choose a repairing response.

Now, the temporal and prosodic signals of the speakers' integration are largely *spontaneous*. Of course, there are cases where a speaker deliberately produces an echoic response with particular prosody in a particular timing, but that must be exceptional, just as a deliberate expression of anger with a face color is exceptional. If so, those temporal and prosodic signals bring in another factor that can break the autonomy of the grounding acts performed with display utterances: even if a speaker ever intends to perform an acknowledgment with an echoic response, its timing and prosody may reveal the low integration on the speaker's part and thus make the utterance function as a repair-request; of course, the opposite is also possible. In either case, the kind of grounding act to be performed by an echoic response is not under the speaker's control.

4. Toward a Dynamic Model for Display Acts

One of the main motivations for Traum (1994) to have developed his own model of grounding is that in the Clark-Schaefer model, it is hard to determine the grounding function of a given utterance "on line," without having to refer to a subsequent development of the dialogue. However, we have just found that the grounding act performed with a display utterance, may it be an acknowledgment or a repair-request, is non-autonomous in two counts. Thus, one may be tempted to think that it cannot be determined on line either. In this view, the grounding act performed by B's response in (1) is fixed only after A responds to it: if A's response is "Yeah, Chestnut Street" (a restatement), it *makes* B's utterance a repair-initiation, whereas if A's response is "And then" (an initiation of new information), it *makes* B's response an acknowledgment.

In the following, we describe a model of display utterances that does justice to the non-autonomy of their grounding functions without being committed to this radical jointness view. We propose redefining the conditions for the grounding acts of acknowledgment and repair-request as follows:

Acknowledgment: In response to A's utterance u, B gives sufficient evidence e that she has correctly identified what A meant by u.

Repair-request: In response to A's utterance u, B gives sufficient evidence e that she has failed to identify what A meant by u.

In both cases, A may not be aware of her identification or lack thereof.

It is helpful to take a layered picture on acts to develop a model that captures the non-autonomous nature of grounding acts. An utterance by A of a certain expression such as "uh huh," "yeah," or "ok" counts as an acknowledgment act by A towards a proposition p under a certain context, e.g., when it is produced as a response to B's utterance of p. This "counts as" relation between acts can be captured by Goldman's *action generation* relation (Goldman, 1970). An act α is said to generate another act β under an appropriate contextual condition C. The same act α, however, may generate a different act β' under a different contextual condition C'. Even though α is in complete control of A, it can generate another act β, which, under a certain context, might not be within her scope of intention.

We characterize a display utterance by the following three components:

(a) the *target* of construal: the dialogue object the display is directed at,

(b) the *content* of construal: what is being displayed, and

(c) the *result* of construal: what cognitive and emotional state one is in.

For the echoic response in (1), the target of B's echoic response is A's utterance of "on Chestnut Street" in the preceding turn, and the content is B's response itself. The result is the integration rate indicated by the prosody that accompanies B's echoic response. Here, the target and the content can be their surface phonological sequences or they can be the semantic contents they express.

The picture we are proposing is (1) a display act α is a lower level autonomous act, characterized by the content and the result components of a display utterance, (2) a display act α generates a grounding act β in a context characterized by the target component, and (3) which grounding

Generating Act (α) Content	Result	Context Target	Generated Act (β)
"uh huh"		following p	acknowledgment
"what?"		following p	repair request
display p	High	following p	acknowledgment
display p'	High	following p	repair
display p	Neutral	following p	acknowledgment
display p'	Neutral	following p	repair request
display p	Low	following p	repair request
display p'	Low	following p	repair request

Table 1: Grounding acts generated by echoic responses.

act is generated is determined instantaneously, but the generated act can be non-autonomous. In the case of echoic responses, a locutionary act of producing an echoic response, characterized by its content and result, generates a certain grounding act under a certain contextual condition, characterized by the target of the echoic response. Table 1 summarizes the types of grounding acts generated out of echoic responses depending on these three parameters. Utterances of typical grounding-oriented expressions are also included in the table for comparison.

In the case of "uh huh" produced after an utterance of p by a partner, it is guaranteed, by the linguistic convention of English, that it serves as evidence that the speaker of "uh huh" believes that she has identified what the previous speaker meant, namely p. Unless further evidence to the contrary is available, it also serves as evidence that the speaker actually identified p. Since this satisfies the condition of an act of acknowledgment, the locutionary act of producing "uh huh" generates an act of acknowledgment. Note that the act of acknowledgment here is both autonomous and instantaneous. Similarly, the locutionary act of producing "what?" generates an autonomous and instantaneous act of repair-request because of the linguistic convention associated with the expression.

An echoic response with high integration prosody provides, if it is a correct echo, two independent pieces of evidence for positive identification, and generates an act of acknowledgment. If it is an incorrect echo, it provides incoherent pieces of evidence. The negative evidence directly obtained from the content defeats the positive evidence, and it becomes an act of repair. An echoic response with neutral prosody, on the other hand, lacks one source of evidence and provides either positive or negative evidence for the identification of the target in the previous turn, and, hence, can generate an act of acknowledgment or an act of repair-request. An echoic response with low integration prosody provides negative evidence for identification, and generates an act of repair-request [1].

The choice of giving up on autonomy in favor of instantaneity for grounding acts might find its support when we consider our daily face-to-face conversations. Most non-linguistic signals are spontaneous displays of one's cognitive and emotional states, which are out of the intentional control of the speakers; people nonetheless invariably exploit these signals to navigate through the course of a conversation. Prosody, which we have found to signal the speaker's integration level in echoic responses, is normally a spontaneous feature of speech. The fact that prosody plays a significant role in grounding by itself shows that grounding acts are non-autonomous.

[1]There appears to be asymmetry between direct evidence and indirect evidence. Indirect evidence provided by prosody generally takes precedence over direct evidence, and only negative direct evidence overrides indirect positive evidence, but not vice versa. This is probably caused by the difference between two types of negativities, the lack of identification and misidentification.

5. Conclusions

In this paper, we argued that the kind of grounding act being performed by a display utterance depends on two major contextual factors, namely, the correctness of the displayed construal and the degree of the speaker's integration as signaled by the characteristics of her speech. Since both factors are generally beyond the speaker's control, we concluded that the acts of acknowledgment and repair-request, when performed with display utterances, are not autonomous acts.

We then proposed a generation model of the grounding functions of display utterances, which sets up multiple layers of generated grounding acts for a single display utterance. In this model, whether an utterance performs an autonomous grounding act depends on which layer of grounding act the question is directed at.

Under our conceptions of the acts of acknowledgment or repair-request, however, whether a display utterance performs one of these acts is fixed at the time of the utterance, without reference to the subsequent development of the dialogue. In other words, the acts of acknowledgments and repair-requests performed by display utterances retain instantaneity without being autonomous. Thus, this paper suggests a way of reconciling the non-autonomous nature of display utterances (Clark, 1996) with the necessity for on-line classifications of the grounding acts (Traum, 1994).

6. References

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13:259–294.

Herbert H. Clark, 1996. *Using Language*. Cambridge University Press.

Alvin I. Goldman. 1970. *A theory of human action*. Princeton University Press, Princeton, NJ.

Atsushi Shimojima, Hanae Koiso, Marc Swerts, and Yasuhiro Katagiri. 1998. An informational analysis of echoic responses in dialogue. In *Proceedings of the 20th Annual Conference of the Cognitive Science Society*, pages 951–956.

Atsushi Shimojima, Yasuhiro Katagiri, Hanae Koiso, and Marc Swerts. 1999. An experimental study on the informational and grounding functions of prosodic features of Japanese echoic responses. In *Proceedings of the ESCA Workshop on Dialogue and Prosody*, pages 187–192.

David R. Traum. 1994. *A Computational Theory of Grounding in Natural Language Conversation*. Ph.D. thesis, University of Rochester.