Lund University, Centre for Languages & Literature, Dept. of Linguistics & Phonetics 117
Working Papers 52 (2006), 117–120

Prosody and Grounding in Dialog

Gabriel Skantze, David House, and Jens Edlund
Department of Speech, Music and Hearing, KTH, Stockholm
{gabriel|davidh|edlund}@speech.kth.se

Abstract
In a previous study we demonstrated that subjects could use prosodic features (primarily peak height and alignment) to make different interpretations of synthesized fragmentary grounding utterances. In the present study we test the hypothesis that subjects also change their behavior accordingly in a human-computer dialog setting. We report on an experiment in which subjects participate in a color-naming task in a Wizard-of-Oz controlled human-computer dialog in Swedish. The results show that two annotators were able to categorize the subjects' responses based on pragmatic meaning. Moreover, the subjects' response times differed significantly, depending on the prosodic features of the grounding fragment spoken by the system.

1 Introduction
Detecting and recovering from errors is an important issue for spoken dialog systems, and a common technique for this is verification. However, verifications are often perceived as tedious and unnatural when they are constructed as full propositions verifying the complete user utterance. In contrast, humans often use fragmentary, elliptical constructions such as in the following example: "Further ahead on the right I see a red building." "Red?" (see e.g. Clark, 1996).

In a previous experiment, the effects of prosodic features on the interpretation of such fragmentary grounding utterances were investigated (Edlund et al., 2005). Using a listener test paradigm, subjects were asked to listen to short dialog fragments in Swedish where the computer replies after a user turn with a one-word verification, and to judge what was actually intended by the computer by choosing between the paraphrases shown in Table 1.

Table 1. Prototype stimuli found in the previous experiment.

Position	**Height**	**Paraphrase**	**Class**
Early	Low	Ok, red	ACCEPT
Mid	High	Do you really mean red?	CLARIFYUNDERSTANDING
Late	High	Did you say red?	CLARIFYPERCEIVE

The results showed that an early, low F_0 peak signals acceptance (display of understanding), that a late, high peak is perceived as a request for clarification of what was said, and that a mid, high peak is perceived as a request for clarification of the meaning of what was said. The results are summarized in Table 1 and demonstrate the relationship between prosodic realization and the three different readings. In the present study, we want to test the hypothesis that users of spoken dialog systems not only perceive the differences in prosody of synthesized fragmentary grounding utterances, and their associated pragmatic meaning, but that they also change their behavior accordingly in a human-computer dialog setting.

2 Method

To test our hypothesis, an experiment was designed in which 10 subjects were given the task of classifying colors in a dialog with a computer. They were told that the computer needed the subject's assistance to build a coherent model of the subject's perception of colors, and that this was done by having the subject choose among pairs of the colors green, red, blue and yellow when shown various nuances of colors in-between (e.g. purple, turquoise, orange and chartreuse). They were also told that the computer may sometimes be confused by the chosen color or disagree. The experiment used a Wizard-of-Oz set-up: a person sitting in another room – the Wizard – listened to the audio from a close talking microphone. The Wizard fed the system the colors spoken by the subjects, as well as giving a go-ahead signal to the system whenever a system response was appropriate. The subjects were informed about the Wizard setup immediately after the experiment, but not before. A typical dialog is shown in Table 2.

Table 2. A typical dialog fragment from the experiment (translated from Swedish).

S1-1a	[presents purple flanked by red and blue]
S1-1b	what color is this
U1-1	red
S1-2	red (ACCEPT/CLARIFYUND/CLARIFYPERC) *or* mm (ACKNOWLEDGE)
U1-2	mm
S1-3	okay
S2-1a	[presents orange flanked by red and yellow]
S2-1b	and this
U2-1	yellow perhaps
[…]	

The Wizard had no control over what utterance the system would present next. Instead, this was chosen by the system depending on the context, just as it would be in a system without a Wizard. The grounding fragments (S1-2 in Table 2) came in four flavors: a repetition of the color with one of the three intonations described in Table 1 (ACCEPT, CLARIFYUND or CLARIFYPERC) or a simple acknowledgement consisting of a synthesized /m/ or /a/ (ACKNOWLEDGE) (Wallers et al., 2006). The system picked these at random so that for every eight colors, each grounding fragment appeared twice.

All system utterances were synthesized using the same voice as the experiment stimuli (Filipsson & Bruce, 1997). Their prosody was hand-tuned before synthesis in order to raise the subjects' expectations of the computer's conversational capabilities as much as possible. Each of the non-stimuli responses was available in a number of varieties, and the system picked from these at random. In general, the system was very responsive, with virtually no delays caused by processing.

3 Results

The recorded conversations were automatically segmented into utterances based on the logged timings of the system utterances. User utterances were then defined as the gaps in-between these. Out of ten subjects, two did not respond at all to any of the grounding utterances. For the other eight, responses were given in 243 out of 294 possible places. Since the object of our analysis was the subjects' responses, two subjects in their entirety and 51 silent responses distributed over the remaining eight subjects were automatically excluded from analysis.

User responses to fragmentary grounding utterances from the system were annotated with one of the labels ACKNOWLEDGE, ACCEPT, CLARIFYUND or CLARIFYPERC, reflecting the preceding utterance type.

In almost all cases subjects simply acknowledged the system utterance with a brief "yes" or "mm" as the example U1-2 in Table 2. However, we felt that there were some differences in the way these responses were realized. To find out whether these differences were dependent on the preceding system utterance type, the user responses were cut out and labeled by two annotators. To aid the annotation, three full paraphrases of the preceding system utterance, according to Table 1, were recorded. The annotators could listen to each of the user responses concatenated with the paraphrases, and select the resulting dialog fragment that sounded most plausible, or decide that it was impossible to choose one of them. The result is a categorization showing what system utterance the annotators found to be the most plausible to precede the annotated subject response. The task is inherently difficult – sometimes the necessary information simply is not present in the subjects' responses – and the annotators only agreed on a most plausible response in about 50% of the cases. The percentage of preceding system utterance types for the classifications on which the annotators agreed is shown in Figure 1.

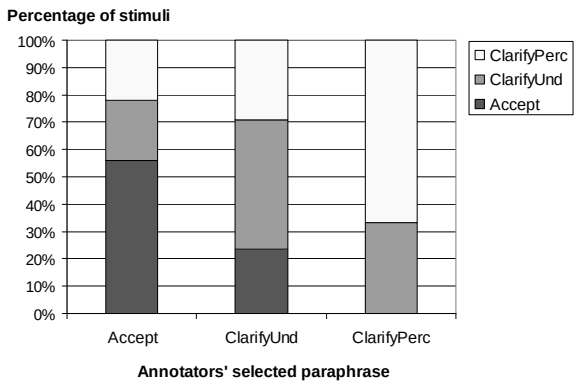

Figure 1. The percentage of preceding system utterance types for the classifications on which the annotators agreed.

Table 3. Average of subjects' mean response times after grounding fragments.

Grounding fragment	Response time
ACCEPT	591 ms
CLARIFYUND	976 ms
CLARIFYPERC	634 ms

Figure 1 shows that responses to ACCEPT fragments are significantly more common in the group of stimuli for which the annotators had agreed on the ACCEPT paraphrase. In the same way, CLARIFYUND, and CLARIFYPERC responses are significantly overrepresented in their respective classification groups (χ^2=19.51; dF=4; p<0.001). This shows that the users' responses are somehow affected by the prosody of the preceding fragmentary grounding utterance, in line with our hypothesis.

The annotators felt that the most important cue for their classifications was the user response time after the paraphrase. For example, a long pause after the question "did you say red?" sounds implausible, but not after "do you really mean red?". To test whether the response times were in fact affected by the type of preceding fragment, the time between the end of each system grounding fragment and the user response (in the cases where there was a user response) was automatically determined using /nailon/ (Edlund & Heldner, 2005), a software package for extraction of prosodic and other features from speech. Silence/speech detection in /nailon/ is based on a fairly simplistic threshold algorithm, and for our purposes, a preset threshold based on the average background noise in the room where the experiment took place was deemed sufficient. The results are shown in Table 3. The table shows that, just in line with the annotators' intuitions, ACCEPT fragments are followed by the shortest re-

sponse times, CLARIFYUND the longest, and CLARIFYPERC between these. The differences are statistically significant (one-way within-subjects ANOVA; $F=7.558$; $dF=2$; $p<0.05$).

4 Conclusions and discussion

In the present study, we have shown that users of spoken dialog systems not only perceive the differences in prosody of synthesized fragmentary grounding utterances, and their associated pragmatic meaning, but that they also change their behavior accordingly in a human-computer dialog setting. The results show that two annotators were able to categorize the subjects' responses based on pragmatic meaning. Moreover, the subjects' response times differed significantly, depending on the prosodic features of the grounding fragment spoken by the system.

The response time differences found in the data are consistent with a cognitive load perspective that could be applied to the fragment meanings ACCEPT, CLARIFYPERC and CLARIFYUND. To simply acknowledge an acceptance should be the easiest, and it should be nearly as easy, but not quite, for users to confirm what they have actually said. It should take more time to reevaluate a decision and insist on the truth value of the utterance after CLARIFIYUND. This relationship is nicely reflected in the data.

Although we have not quantified other prosodic differences in the users' responses, the annotators felt that there were subtle differences in e.g. pitch range and intensity which may function as signals of certainty following CLARIFYPERC and signals of insistence or uncertainty following CLARIFYUND. More neutral, unmarked prosody seemed to follow ACCEPT. When listening to the resulting dialogs as a whole, the impression is that of a natural dialog flow with appropriate timing of responses, feedback and turntaking. To be able to create spoken dialog systems capable of this kind of dialog flow, we must be able to both produce and recognize fragmentary grounding utterances and their responses. Further work using more complex fragments and more work on analyzing the prosody of user responses is needed.

Acknowledgements

This research was supported by VINNOVA and the EU project CHIL (IP506909).

References

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

Edlund, J. & M. Heldner, 2005. Exploring Prosody in Interaction Control. *Phonetica 62*(2-4), 215-226.

Edlund, J., D. House & G. Skantze, 2005. The effects of prosodic features on the interpretation of clarification ellipses. *Proceedings of Interspeech 2005*, Lisbon, 2389-2392.

Filipsson, M. & G. Bruce, 1997. LUKAS – a preliminary report on a new Swedish speech synthesis. *Working Papers 46*, Department of Linguistics and Phonetics, Lund University.

Wallers, Å., J. Edlund & G. Skantze, 2006. The effect of prosodic features on the interpretation of synthesised backchannels. *Proceedings of Perception and Interactive Technologies*, Kloster Irsee, Germany.