

THE EFFECT OF CONTEXT ON THE INTELLIGIBILITY OF DIALOGUE

David G. Novick, Karen Ward & Benjamin Corliss
email: novick@cse.ogi.edu, wardk@cse.ogi.edu
Center for Spoken Language Understanding
Oregon Graduate Institute of Science & Technology
P.O. Box 91000
Portland, Oregon 97291-1000
USA

ABSTRACT

We measured the effects of task context on the intelligibility of utterances drawn from a corpus of air traffic control dialogue. Subjects understood more words when the utterances were presented in the form of a dialogue, with the original utterance order preserved. Subjects presented with the same utterances in randomized order understood significantly fewer words. This effect was seen with both domain experts (pilots and air traffic controllers) and domain novices.

1. INTRODUCTION

To improve performance, spoken language understanding systems rely on constraints derived from knowledge of the task they are designed to perform. Most speech recognizers assume a task-specific vocabulary, for example, and many systems use task-specific grammars. To improve system performance, different grammars may be used to constrain speech recognition during different portions of the interaction, reflecting the expectation that the user is likely to say different things at different points in the task. The Minds-II system [9] dynamically generates a recognizer grammar based on context, inferred plans, and goals. The automated scheduling system described in [3] maintains contexts for interpreting users' utterances and specializes these contexts in a stack-based fashion. These contexts are then used to predict the utterance grammars for both recognizer and parser.

The fundamental potential of this technique is unknown, however, particularly for mixed-initiative dialogue. Does the context of an utterance that is part of a complex task contribute to its intelligibility?

We address the question of context and intelligibility by examining the extent to which task context aids human listeners in understanding spoken language. By "task context" we mean the state of the underlying domain activity that forms the reason for the dialogue. If task context plays a role in intelligibility, then the effect of context (or its absence) should be demonstrable in a task-focused dialogue. In this study, we presented subjects with utterances recorded from such a dialogue. Utterances were presented either in the order in which they had been produced in the original conversation or in a

randomized order. By measuring the numbers of words that subjects were able to understand, we hoped to gain some insight into the contribution that task context makes to intelligibility.

That sentence context can affect word recognition is well-documented [7]. Such studies have typically focused on examining the effects of semantic priming or phonetic context on the interpretation of an acoustically obscured or ambiguous word within an otherwise-clear presentation (e.g., [2]). Furthermore, Schober and Clark [6] have demonstrated the importance of dialogue-level effects in relating complex descriptions to abstract pictures. In this study, however, we were interested in examining the effect of dialogue-level context on the hearer's ability to understand the words in a sentence as a whole.

In this paper we describe an experiment designed to test the hypothesis that utterances presented in the context of a task-oriented dialogue would be more intelligible than the same utterances presented in a decontextualized sequence. We first describe the task and domain from which the test utterances were drawn. We then describe the experiment and discuss our conclusions.

2. TASK DOMAIN

Under normal circumstances, human listeners are so proficient at understanding even decontextualized speech that it can be difficult to detect the contribution that task context provides. We therefore turned to a particularly difficult communications task, that of understanding air traffic control (ATC) dialogue.

In this study, we presented both domain experts (pilots and air traffic controllers) and domain novices (non-pilots) with recorded utterances drawn from a corpus of ATC dialogue [8] and asked them to repeat the words they heard. This proved to be a task that was difficult enough to reveal task context effects.

2.1. Characteristics of ATC Dialogue

Air Traffic Control communication takes place over noisy, low-bandwidth radio channels. The utterances contain many letters and numbers, a notoriously confusable vocabulary. Furthermore, the participants are strongly encouraged to use a constrained vocabulary and phrase

- (112) Controller: Delta fourteen forty three turn left heading three one zero join the localizer.
 (113) Pilot: Delta fourteen forty three, three one zero join the local-
Controller speaks to another aircraft. Several minutes later:
- (121) Controller: Delta fourteen forty three you've got niner miles to Laker maintain three thousand till established on the localizer cleared I L S two eight right approach maintain one seven zero knots until Laker.
 (122) Pilot: Delta fourteen forty three, cleared for the I L S to two eight right and uh, maintain three thousand, till on the localizer.
 (123) Controller: Delta fourteen forty three that's correct and maintain one seven zero knots until Laker.
 (124) Pilot: Copy one seven zero knots till Laker Delta fourteen forty three, thank you.

Figure 1: Excerpts from an air traffic control conversation

This excerpt includes two exchanges. The first exchange, utterances 112-113, illustrates a simple instruction-readback pair. In the second exchange, utterances 121-124, the pilot read back only part of the instructions and so a clarification subdialogue ensued. Utterance numbers are from the original transcript [8].

structure [4, 5]. The purpose of these constraints is to improve the intelligibility of safety-critical communications but the overall effect can be to make the utterances even less understandable to a domain novice. Also, ATC dialogue is typically delivered at a very fast speaking rate, further decreasing its intelligibility.

The structure of ATC dialogue is also unusual. A pilot-controller conversation consists of a collection of brief exchanges, often many minutes apart and interleaved with conversations between the controller and other pilots. To ensure that information and instructions were conveyed successfully, the hearer will repeat the substance of the speaker's statement in a confirming repetition termed a "readback." This leads to a distinctive two-utterance exchange in which the second transmission largely duplicates the content of the preceding one (Figure 1).

Although the conversational context of an individual exchange may be accessible even to a novice, especially in the case of a simple instruction-readback pair, the task context between exchanges is less obvious and requires some knowledge of aviation to discern. This characteristic allowed us to probe the importance of deep task context.

2.2. The Approach Task

Each of the three dialogues used in this study represents an actual, complete conversation between an air traffic controller and the pilot of a commercial flight approaching the airport to land. Controller and pilot are cooperatively performing the task of guiding the aircraft through the controller's airspace to a point from which the pilot can complete the approach and landing without further guidance from the controller. This task is referred to as an Instrument Landing System (ILS) approach procedure [4, 5].

3. EXPERIMENT

We hypothesized that utterances in their task context would be more intelligible than the same utterances presented in a decontextualized sequence. As a validation measure, we predicted that the experts should have higher overall intelligibility scores than the novices. We measured the effects of task context on intelligibility using a 2x2 between-subjects design. The independent variables were (a) availability of task context and (b) domain expertise. The dependent measure was intelligibility, as measured by the number of words that the subjects were able to repeat following the end of the utterance.

The 24 subjects in this study consisted of 12 domain experts and 12 domain novices. The domain experts were pilots or air traffic controllers with experience levels ranging from a private pilot with 80 hours of flying experience to a commercial pilot with over 15,000 hours of flying experience. The domain novices had no prior exposure to air traffic control communications, with one exception: one had several hours' experience as a passenger in small aircraft. Excluding his data from the analysis did not change the results, however, so we retained those data in our final analysis. Most of the domain novices were staff or students at Oregon Graduate Institute. All subjects were unpaid volunteers.

3.1. Method

Task context was manipulated by presenting the ATC utterances in their original task order or in a randomized order. Subjects listened to three complete ATC approach-task dialogues, one utterance (transmission) at a time. The first of the dialogues was a practice case for familiarization with the experimental task; it was not coded nor was it considered in the experimental analysis. In the practice dialogue, the utterances were presented in order. The utterances of the second and third dialogues (the test corpus) were presented either (a) in order or (b) in a ran-

Table 1: Mean Words Correct/Utterance

Test	Subject group	Total Words Correct			Net Words Correct		
		Ordered	Unordered	p value	Ordered	Unordered	p value
All utterances	Novice	4.3	3.7	0.003	-3.3	-4.5	0.005
	Expert	8.6	6.7	< 0.001	4.7	1.1	< 0.001
Readbacks only	Novice	4.7	3.7	0.002	-1.2	-3.0	0.003
	Expert	7.8	6.5	< 0.001	4.7	2.1	< 0.001
Non-readbacks only	Novice	4.1	3.7	0.159	-4.7	-5.6	0.163
	Expert	9.1	6.9	0.014	4.8	0.5	0.015

domly permuted order consistent across the subjects in this condition.

The test corpus comprised a total of 33 utterances with approximately 3400 words. Utterance length ranged from 3 to 31 words, with the average length being 11.4 words. There were 299 words in the vocabulary, with an average branching factor of 3.953. For the purposes of this study, an utterance was defined as being a single transmission. None of the dialogues included any overlapped speech. All were complete conversations, in that they included all exchanges between pilot and controller from the time that the pilot established contact with the controller until the time that the controller authorized the pilot to contact the next controller.

Utterances were audio-taped from radio, digitized, and presented to subjects over headphones. In debriefing, the pilots judged the signal quality to be typical of that encountered in actual flying conditions.

Subjects sat at a computer console and pressed a key to play each utterance. Each utterance was presented only once. They were instructed to repeat the utterance immediately, word for word and as accurately as possible, and this verbal response was recorded. They were then asked to type the words they had just uttered. Where the verbal and the written responses differed, we took the subject's verbal response as the authoritative answer.

We used this verbal-plus-written protocol because we were concerned that the verbal responses might include acoustically ambiguous or difficult-to-understand words. We did not want to rely on the typed answers, however, because some of the utterances are fairly long and we thought it likely that subjects would forget what they had heard before they could finish typing. Although there were cases where the written response was more accurate than the spoken response, we based our analysis solely on the spoken responses.

The results were scored in two ways. First, because our primary measure was the number of words understood,

we counted the total number of words correct regardless of word order or insertions. Second, we measured net response accuracy, taking account of word order and penalizing insertions, deletions, and substitutions equally. Some subjects included meta-responses in their typed or verbal response, e.g., "?????", "I can't type, either." These meta-responses were not included in scoring. Responses were scored by two raters, with no disagreements found.

3.2. Results

We compared the average scores for each utterance across conditions using a two-tailed paired t-test. Results are summarized in Table 1. We found that context plays a significant role in intelligibility. In the expert group, we found that the subjects in the ordered condition performed significantly better than subjects in the unordered condition. As expected, the non-expert group found the utterances significantly less intelligible than the experts in both the ordered and the unordered conditions ($p < 0.001$), which validated our primary measure.

The sample of experts was not large enough to test for a correlation between level of aviation experience and score.

The extent to which these results can be attributable to context can be seen by comparing the main ordered/unordered effect across cases where there was not an immediate readback of a prior utterance. Recall that in the ATC domain the second utterance in many two-utterance exchanges is a readback which largely duplicates the preceding utterance. For readbacks, we would expect to see context effects for both experts and non-experts; where the utterance was not a "readback" only the expert group should have access to task context. The data confirm this. Subjects in the expert group had significant context effects for utterances both in the 13 "readback" cases and in the 20 "non-readback" cases. Subjects in the non-

expert group had a significant context effect in the “read-back” case but not in the “non-readback” case.

As may be expected, we found a correlation (0.5825 , $r^2=0.3393$, $p < 0.001$) between brevity of utterance and word scores; the shorter the utterance the more likely the subject was to report the words correctly.

4. CONCLUSIONS AND FUTURE WORK

The results suggest that for mixed-initiative dialogue, task context plays a significant role in understanding spoken language. On average among the subjects we observed, adding context led to a 15 percent improvement in intelligibility for experts in the ATC approach task. While we are confident that this difference is a real one, we are less confident about the magnitude of the difference. We suspect that in a less demanding communications task we would see a smaller task context effect.

In this study, we confirmed system-builders’ intuitions that domain dialogue context plays a role in word recognition. Even human domain experts, who have a strong domain vocabulary model, performed better when dialogue context was available. From this we conclude that domain dialogue models are needed in spoken language understanding systems not only for recognition purposes, as they are now, but also to better model user understanding of system contributions.

ACKNOWLEDGEMENTS

The authors thank Brian Hansen for his helpful comments. This research was supported by the National Science Foundation.

REFERENCES

- [1] R. Cole, D. Novick, M. Fanty, S. Sutton, B. Hansen, & D. Burnett. “Rapid Prototyping of Spoken Language Systems: The Year 2000 Census Project,” *Proceedings of the International Symposium on Spoken Dialogue (ISSD-93)*, Tokyo, November, 1993, pp. 19-23.
- [2] C. Connine. “Effects of Sentence Context and Lexical Knowledge in Speech Processing,” *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*, Gerry T. M. Altman (ed.), MIT Press, Cambridge, Mass., 1990, pp. 281-294.
- [3] Fanty, M., Sutton, S., Novick, D., & Cole, R. “Automated Appointment Scheduling,” *ESCA Workshop on Spoken Dialogue Systems*, Vigsø, Denmark, May, 1995, pp. 141-144.
- [4] Federal Aviation Administration. *Air Traffic Control*, 7110.65F, 1989.
- [5] Federal Aviation Administration. *Airman’s Information Manual*, 1991.
- [6] Schober, M. F. & Clark, H. H. “Understanding by Addressees and Overhearers,” *Cognitive Psychology* (21), 1989, pp. 211-232.

[7] L. Tyler. “Context and Sensory Input,” *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives*, Gerry T. M. Altman (ed.), MIT Press, Cambridge, Mass. 1990, pp. 315-323.

[8] K. Ward, D. G. Novick, & C. Sousa. “Air Traffic Control Communications at Portland International Airport,” Technical Report No. CS/E 90-025, Oregon Graduate Institute, Beaverton, OR, 1990.

[9] S. Young & W. Ward. “Semantic and Pragmatically Based Re-recognition of Spontaneous Speech,” *Eurospeech '93*, 1993, pp. 2243-2246.