Predictable and Educational Spoken Dialogues: Pilot Results

Gregory Aist[1] *and Jack Mostow*[2]

[1]Language Technologies Institute and [2]Robotics Institute
Carnegie Mellon University, Pittsburgh, USA
Gregory.Aist@alumni.cmu.edu

Abstract

This paper addresses the challenge of designing spoken dialogues that are of educational benefit within the context of an intelligent tutoring system, yet predictable enough to facilitate automatic speech recognition and subsequent processing. We introduce a design principle to meet this goal: construct short dialogues in which the desired student utterances are external evidence of performance or learning in the domain, and in which those target utterances can be expressed as a well-defined set. The key to this principle is to teach the human learner a process that maps inputs to responses. Pilot results in two domains – self-generated questions and morphology exercises – indicate that the approach is promising in terms of its habitability and the predictability of the utterances elicited. We describe the results and sketch a brief taxonomy classifying the elicited utterances according to whether they evidence student performance or learning, whether they are amenable to automatic processing, and whether they support or call into question the hypothesis that such dialogues can elicit spoken utterances that are both educational and predictable.

1. Introduction

When designing spoken dialogue, researchers and practitioners often attempt to elicit predictable speech from users. The reason is clear: speech recognition is hard for computers, and especially difficult for children's speech [1, 2, 3]. Thus dialogues are often designed to increase predictability and thus decrease difficulty for the speech recognizer.

Intelligent tutoring systems, on the other hand, are designed to yield interactions which are of educational benefit to the student. Sometimes this design constraint means that the interaction takes longer or requires more work on the part of the student than would be the case if the interaction were designed for maximum efficiency of task performance: for example, rather than giving the student an answer, the software may provide a hint.

These twin challenges – predictability on the one hand, and educational effectiveness on the other – are often at odds. If dialogue is designed to be maximally predictable, adopting strategies directly from task-oriented dialogue might yield situations such as in Figure 1, where asking the student to choose an answer from a spoken list inadvertently transforms the student's task from recall into recognition, which may not be as educationally beneficial.

> Show: Joe can read very well.
> Ask: Do you say this like reed or like red?
> Listen for: { *read*/*reed*, *read*/*red* }

Figure 1: *Spoken multiple choice, a common technique in task-oriented dialogue, applied to word reading. The technique requires giving users a list of available options, and thus gives away the answer.*

A number of tutorial dialogue systems have successfully navigated between the twin challenges of predictability and effectiveness. Pronunciation practice using speech recognition has a long history (for reviews, see [4, 5]). A common pronunciation exercise is for the software to present the learner with a single word, a phrase, or a sentence in a second language, listen for the pronunciation, and respond appropriately. These exercises often aim for predictability through displaying on screen text for the learner to speak aloud, or through constructing the conversational setting to produce strong expectations about what might be appropriate to say next. Automated pronunciation tutoring has been demonstrated to be effective for targeted sounds (e.g. [6]) and in at least some real-world situations (e.g. [7]). In the domain of reading, Project LISTEN's Reading Tutor uses speech recognition to listen to children read aloud, and helps them learn to read. The Reading Tutor implements the instructional paradigm of guided oral reading, and is designed for predictability by displaying on screen the text for the student (rather than, say, trying to follow along on a printed book) and by listening for at most one sentence at a time. The Reading Tutor's effectiveness has been tested in an extensive series of empirical evaluations [8, 9].

We introduce in this paper the following design principle aimed at meeting the twin goals of educational utility and linguistic predictability: Design dialogues in which the target utterances show task learning or performance, and form a well-defined, predictable set. We operationalize this strategy as follows: Elicit a predictable response by teaching a process for mapping prompt to response; a student who learns that process will predictably produce the response when given the prompt. The process is parameterized, which allows for its applicability to tasks beyond paired associate learning, as we show in this paper by applying the strategy to comprehension questions and to morphology exercises.

In the case of comprehension questions, the idea is to teach a process for generating a question: pick a character, pick a wh-word, and pick an action; then, compose and speak a question such as *WH- did CHARACTER ACTION?* Multiple correct responses are possible, yet the desired outcome set is constrained enough to be predictable.

In the case of the morphology dialogue, the idea is to teach a parameterized collaborative process to infer word meaning: obtain the meaning of the prefix, obtain the meaning of the stem; then, compose and speak a gloss of the word combining the meanings of the prefix and the stem. For example: *re- here means again; what does reinvent mean?* (listen for) *invent again*. Here, one correct response is possible, and producing it is evidence of processing the prefix and the word as desired. A later use of this dialogue could rely on the student to supply the prefix meaning, making the student do more work when educationally appropriate.

2. Design of spoken dialogue: comprehension

As part of current research on automatic tutoring for children's reading, we set out to design spoken dialogue to help children learn comprehension strategies, specifically for self-questioning. Previously Project LISTEN had tried applying spoken dialogue techniques to the analysis of free-form spoken responses to open-ended questions; this turned out to be quite difficult, even to the point where word-spotting when all of the words in the transcript were given to the recognizer still resulted in low accuracy. It became evident that reframing the problem might help address the accuracy issue. But how to increase the predictability of the expected answers, without resorting to shallow questions or menus as in Figure 1? Concurrently, Wei and Mostow were developing methods for using a computational model of mental states to automatically generate questions from narrative text [10], for example:

Why were the mice afraid of the cat?

For a particular text, such questions can be conceived of as a finite language consisting of a fixed grammar portion plus concepts abstracted from the text, in this case a fable including two mice who are eating food and see a cat. The wh-words and similar classes are the fixed portion, and actions and characters from the text are the variable portion. [10] describes automatic methods and current performance of, extraction of actions and characters.

The next step is to take this predictable set of questions and produce a dialogue in which the tutoring system would listen for them, and in which the student's production of them would be educationally valuable. Having children generate questions about the text is known to be educationally beneficial ([11], p. 15) so in this case we simply have to be careful of giving away too much of the question: The system can ask the student to create a spoken question, perhaps specifying parts of it ahead of time (Figure 2).

Underlying process: Select a character and a wh-word and construct a question about something that character did or experienced in the story Say: a prompt to create or complete a question Listen for: a generated question of the expected form

Figure 2: *Dialogue design for generating questions.*

3. Design of spoken dialogue: morphology

We also set out to design spoken dialogue to help children learn vocabulary strategies, specifically strategies for deciphering the meaning of words based on morphology. We wanted to construct dialogues that would illustrate how English prefixes contribute to the meanings of words, to be used opportunistically when students encountered words containing those prefixes during the course of oral reading.

The idea, therefore, was to develop a spoken dialogue intervention that provides relatively short interventions on words so as to avoid disrupting the story flow, is aimed at helping students comprehend the text at hand, and familiarizes students with high-utility morphology, in a developmentally appropriate way.

So, when a difficult and/or complex word is encountered:

1. Identify the stem and/or the core meaning,
2. If it has a reliable morphological cue, illustrate the cue by showing its use in that word (if appropriate) or a simpler word. Figure 3 illustrates the dialogue strategy: show the word used to illustrate the prefix, explain the meaning of the prefix, and prompt for a paraphrase of the word that contains the meaning of the stem and the meaning of the word.

Underlying process: Combine meaning of prefix and meaning of word to produce a short gloss Show: reinvent Say: *reinvent* Short pause Say: *re- here means again* Short pause Say: *what does reinvent mean?* Listen for: *invent again*

Figure 3: *Dialogue design for scaffolding morphology*

These dialogue strategies for questions and morphology have undergone preliminary field tests, which we now describe. In the pilot studies that follow, we categorized each utterance as one of: Empty, Predicted answer, Correct but not predicted, or Incorrect. For the utterances containing the Predicted answer, we also marked whether the utterance was an exact match (Predicted=), was a match up to slight variation in ending as in behave vs. behaved (Predicted~), or was an exact or variational match but contained other material (Predicted+). We use this coding scheme in both the two results sections that follow, as well as the Discussion.

4. Pilot results for comprehension strategies

We field-tested these comprehension strategy exercises with five students on the text shown in part below.

The Country Mouse and the Town Mouse

Once upon a time a town mouse went on a trip to the country. There he met a country mouse. They soon became friends. … So the town mouse invited the country mouse to visit him in the city. …When the country mouse got to town and saw the cheese, cake, honey, jam and other goodies at the house, he was pleasantly surprised. … The town mouse said, "You're my guest, so dig in!" They began to feast, and the country mouse tried to taste everything before his tummy was full. … Suddenly there came the sound of heavy footsteps. The two mice ran. The man of the house had come to get a snack. He saw that mice had gotten some honey. So he decided to send the cat. The mice, full of terror, hid away. They didn't make a sound. …

The computer tutor provided for the students what the strategy was ("Questioning is…"), its importance, examples ("Why is it a pleasant surprise for the country mouse to see all those good snacks?"), and practice forming questions by picking from a menu a character (the town mouse, the country mouse, the man of the house, the cat), a question type (why/who/what), and an action. Students were then given two opportunities to form questions: one, by choosing from a menu the character type and the question and then completing the question verbally; and two, by being prompted for a complete question. We collected the following examples of student questions generated either entirely or completely by speech, as described above, and coded them for correctness and predictedness as previously described. (Only non-null responses are shown here.)

Subj.	*utterance*	label
mKJ	*how did the cat see the mice*	predicted=
mKJ	*why did the two mice come out* (speech to other kid omitted)	predicted=
mAJ	*wh- why did the country mouse leave*	predicted+
fAG	*i would i would like to know about the town mouse and the country mouse being friends*	correct
mDB	*why did the man in the house uh why did the man of the house why did the man of the house try to hurt th- um things the mice*	predicted+
mDG	*did the did the*	incorrect
mDG	*how did the man of the house decide to send the cat how did the man of the house decide to send the cat*	predicted+

5. Pilot results for morphology exercises

We field-tested these morphology exercises with two students who were using the Reading Tutor. In this case, each student received several opportunities to try the exercises, with a range of prefixes and words. Students sometimes restarted the story, so some words were attempted more than once; each student's first recorded attempt at each word is shown in **bold** with subsequent attempts non-bold.

Subject fNH	*utterance*	label
repaint	***again <sil> paint again***	predicted+
misbehave	***badly behaved <sil> behaved***	correct
nonswimmer	***can't swim***	correct
repaint	*paint again*	predicted=
misbehave	*behaved badly*	predicted~
nonswimmer	*not a swimmer*	predicted=
unhappy	***not happy***	predicted=
prepay	***i have ahead of pay***	correct
underfed	***not enough fed***	correct
retell	***tell again or***	predicted+
nonexpert	***not a expert***	predicted~
unable	***not able to do it***	predicted+
repaint	*<sil>*	empty
misbehave	*behave badly*	predicted=
nonswimmer	*can't swim*	correct
unhappy	*not <sil> not happy*	predicted+
prepay	*ahead of time*	incorrect
underfed	*not fed enough*	predicted=
retell	*tell again*	predicted=
mislabel	***badly <sil> uh <sil> labeled badly***	predicted+
nonexpert	*not an expert*	predicted=
unable	*un <sil> unable to do something*	incorrect
preheat	***pay ahead of time***	incorrect
underused	***not enough used <sil> not used enough i think***	predicted+
Subject fAM		
repaint	***repeat repeat***	incorrect
misbehave	***he was being bad in school***	correct
nonswimmer	***nonswimmer means you do not <cutoff>***	incorrect
unhappy	***you're not happy***	predicted+
prepay	***prepay <sil> prepay means i don't know <sil> ahead <sil> oh well i'm not gonna do it***	incorrect
underfed	***underfed means um <sil> un***	incorrect
retell	***retell means you tell again huh <sil> do i push do i push go***	predicted+
mislabel	***oh mislabel mean***	incorrect
nonexpert	***non <sil> not an expert oo hoo***	predicted+
unable	***unable means***	incorrect
preheat	***preheat <sil> preheat means you turn on the heat over and over <sil> over***	incorrect
underused	***unders <sil> un <cutoff>***	incorrect

6. Discussion and Conclusion

First, we categorize the non-empty utterances according to what kind of answer they contain, and give summary statistics. Second, we discuss for each category the expected difficulty of classifying utterances into that category. Third, we discuss whether the pilot data supports or calls into question the hypothesis that these dialogues elicit predictable and educational speech.

For the question generation exercises the counts were:

2 predicted=
0 predicted~
3 predicted+
1 correct
1 incorrect

For morphology the counts for first attempts per word were:

fNH:	fAM:
1 predicted=	0 predicted=
1 predicted~	0 predicted~
5 predicted+	3 predicted+
4 correct	1 correct
1 incorrect	8 incorrect

In question generation, 5 out of the 6 acceptable responses recorded contained a predicted response, and 2 of those 5 were the exact response. In morphology, 10 out of the 15 acceptable responses recorded contained a predicted response, and only 1 out of those 10 contained the exact response. Thus it is clear that a key challenge is being able to recognize responses that contain the predicted response while tolerating disfluencies or partial attempts.

How difficult can we expect each of these categories to be to recognize? Empty utterances can be expected to be sometimes recognized as silence, and at other times recognized as short "filler" words. For comprehension questions empty utterances should thus not pose a substantial problem, although for morphology exercises they might if there are not enough options in the language model to provide other matches for background noise or other sounds.

Predicted utterances come in three categories. Predicted= can be expected to be the most accurate category of predicted utterances (vs. predicted+ or predicted~) to recognize, as they contain an exact match to the expected response with no other words to throw off the recognizer. Predicted+ will require robust recognition: either treating the unpredicted portions as disfluencies, or trying to spot the predicted portions in the whole utterance. Predicted~ will require flexibility in terms of the exact wording of the response: the language model might need to include both the exact word (*behave*) and morphological variants (*behaved*), or the recognition process might need to (mis)recognize variants as the original word – a process which often inadvertently occurs anyway with automatic speech recognition, especially if the language model contains only the original word.

Correct but not predicted utterances will presumably be the most difficult to recognize automatically, as they can range from simple rephrasings (*can't swim*) to more complicated expressions (*he was being bad in school*). While we didn't find any instances in this pilot study, correct but not predicted utterances could even include answers that were semantically valid yet contained none of the content words in the stimulus or in the predicted response, such as *preheat* meaning *turn on the oven to get ready to cook*. Philosophically two approaches to correct but not predicted answers are possible: one is to treat them as correct answers that the algorithm should be extended to handle or predict, and the other is to treat them as incorrect answers since they don't follow the "rules" of the language game that the tutor is trying to teach the student. (This is reminiscent of game shows where the answer is required to be in a particular syntactic format, such as a question on the quiz show *Jeopardy*.) For the time being we remain neutral on this issue.

Finally of course if a student gives an incorrect answer, the tutor should recognize that the answer is incorrect, although it likely does not need to be able to automatically transcribe the answer. The difficulty of this process naturally depends on how variable the incorrect answers are and how closely they resemble the predicted and/or correct yet unpredicted answers.

In summary, in this paper we presented a design principle for predictable yet educational spoken dialogue, gave two examples of dialogues in comprehension and morphology, and presented pilot results. Do the pilot data support or call into question the hypothesis that such dialogues elicit predictable and educational speech? The pilot results are promising – most acceptable answers were predicted from the process – yet indicate that while students can and often do produce the target answer, their utterances may also include variants and disfluencies or be acceptable yet not predicted.

7. Acknowledgements

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B070458 and Grant R305A080157. The opinions expressed are those of the authors and do not necessarily represent the views of the Institute and the U.S. Department of Education.

8. References

[1] Russell, M., and D'Arcy, S., "Challenges for computer recognition of children's speech", SLaTE 2007.

[2] Potomanios, A. and Narayanan, S., "A Review of the Acoustic and Linguistic Properties of Children's Speech", *Proceedings of the International Workshop on Multimedia Signal Processing*, October 2007.

[3] Li, Q., and Russell, M. "Why is automatic recognition of children's speech difficult?", EUROSPEECH 2001.

[4] Hincks, R., "Speech Technologies for Pronunciation Feedback and Evaluation", *ReCALL* 15(1):3-20, 2003.

[5] Aist, G., "Speech recognition in computer assisted language learning", In K. C. Cameron (ed.), *Computer Assisted Language Learning (CALL): Media, Design, and Applications*. Lisse: Swets & Zeitlinger, 1999.

[6] Neri, A., Cucchiarini, C., Strik, H. "The effectiveness of computer-based speech corrective feedback for improving segmental quality in Dutch", *ReCALL* 20:225-243, 2008.

[7] Eskenazi, M., Kennedy, A., Ketchum, C., Olszewski, R., and Pelton, G., "The NativeAccent™ pronunciation tutor: measuring success in the real world", SLaTE 2007.

[8] Mostow, J. "Experience from a Reading Tutor that listens: Evaluation purposes, excuses, and methods", in C. K. Kinzer & L. Verhoeven (Eds.), *Interactive Literacy Education: Facilitating Literacy Environments Through Technology*, pp. 117-148. New York: Lawrence Erlbaum Associates, Taylor & Francis. 2008.

[9] Mostow, J., Aist, G., Burkhead, P., Corbett, A., Cuneo, A., Eitelman, S., Huang, C., Junker, B., Sklar, M. B., & Tobin, B. "Evaluation of an automated Reading Tutor that listens: Comparison to human tutoring and classroom instruction". *Journal of Educational Computing Research*, 29(1), 61-117. 2003.

[10] Mostow, J., and Chen, W. "Generating Instruction Automatically for the Reading Strategy of Self-Questioning". AIED 2009.

[11] National Institute of Child Health and Human Development. "Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction." NIH Publication No. 00-4769. Washington, DC: U.S. G.P.O. 2000.