# Designing Spoken Tutorial Dialogue with Children to Elicit Predictable but Educationally Valuable Responses

*Gregory Aist* [1], *Jack Mostow*[1]

[1] Project LISTEN, Carnegie Mellon University, USA

gregory.aist@alumni.cmu.edu, mostow@cs.cmu.edu

## Abstract

How to construct spoken dialogue interactions with children that are educationally effective and technically feasible? To address this challenge, we propose a design principle that constructs short dialogues in which (a) the user's utterance are the external evidence of task performance or learning in the domain, and (b) the target utterances can be expressed as a well-defined set, in some cases even as a finite language (up to a small set of variables which may change from exercise to exercise.) The key approach is to teach the human learner a parameterized process that maps input to response. We describe how the discovery of this design principle came out of analyzing the processes of automated tutoring for reading and pronunciation and designing dialogues to address vocabulary and comprehension, show how it also accurately describes the design of several other language tutoring interactions, and discuss how it could extend to non-language tutoring tasks.

**Index Terms**: spoken dialogue, intelligent tutoring systems.

## 1. Introduction

A key question in speech-enabled intelligent tutoring systems is how to design technically feasible, educationally effective spoken tutorial dialogue. This challenge is especially acute with children, for whom speech recognition is particularly difficult [1, 2]. Recognizing speech is hard for computers, children's speech even harder (even for humans); the feasibility of spoken dialogue with children depends to a great extent on the predictability of speech.

These twin challenges – predictability on the one hand, and educational effectiveness on the other – are often at odds. Naturalistic dialogue, based on human tutor behavior, draws on the effectiveness of human tutoring behaviors – yet automatic speech recognition lacks the accuracy of human speech recognition, especially for children. Directed dialogue, designed for high recognition performance, might trade off predictability and cognitive benefit in perhaps undesirable ways. For example, the common technique from task-oriented dialogue of giving a spoken multiple-choice list, when transferred to tutorial dialogue, requires that the correct answer be present in the list (Figure 1). From the point of view of the student, asking the question in this way changes recall into recognition which may not be as educationally beneficial.

> Show: **dog**
> Ask: Is this word *can*, *dog*, *fog*, or *doll*?
> Listen for: {can, dog, fog, doll}

Figure 1: *Spoken multiple choice, a common technique in task-oriented dialogue, applied to word reading.*

A number of tutorial dialogue systems have successfully navigated between the twin challenges of predictability and effectiveness. We discuss example systems from language tutoring for reading and pronunciation.

**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:

1. Displaying on screen the text for the student to read aloud,
2. Listening for at most one sentence at a time.

The Reading Tutor's effectiveness has been demonstrated in an extensive series of empirical evaluations [3, 4].

**Pronunciation.** Pronunciation practice using speech recognition has a long history [for reviews, see 5, 6]. 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 in one or more of the following ways: by playing back the learner's response together with a reference pronunciation, by showing a graphical representation of the learner's response with a graphical representation of a reference pronunciation, or by analyzing the learner's response and providing specific feedback on what problems were heard, and how to fix them. (Positive feedback may also be provided.) These exercises often aim for predictability through displaying on screen text for the learner to speak aloud. Automated pronunciation training has been demonstrated to be effective for targeted sounds [7, 8] and in at least some real-world situations [9].

## 2. Design of spoken dialogue to teach comprehension strategies

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 strategies 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 list techniques as in Figure 1? Concurrently, Wei and Mostow were developing methods for automatically generating questions from narrative text [under review/to appear], using a computational model of mental states, to model for students the process of building a question to ask yourself. The questions generated have the following general format:

*Why were the mice afraid of the cat?*

As it turns out, the questions based on any particular text passage form a finite language consisting of a fixed grammar

portion plus concepts abstracted from the text (Figure 2), in this case a fable involving two mice who are looking for food and see a cat. This grammar recognizes lots of appropriate questions ("Why did the country mouse fear the cat?") as well as questions which are silly given this particular text ("Where did the city mouse freeze the cat?").

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 [10] 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 the character ahead of time to reduce variability further (Figure 3).

S → Wh Aux Character Action (Character|Item)
S → Wh CopVerb Character State (PP)
PP → P (Character|Item)
P → about | … | under
CopVerb → are | aren't | is | isn't | were | weren't
Wh → what | why | how
Aux → did | didn't | had | hadn't
Character → the cat | the city mouse
| the country mouse | the mice | the mouse
Action → go | go back | fear
| forget | forgot | freeze | see
Item → his own food | the house | town
State → afraid | happy | hungry | ... | sad

Examples: *Why did the mouse freeze?*
*Why was the country mouse afraid?*

Figure 2: *Finite language describing wh-questions generated from text. Closed class nonterminals are mapped to a fixed set of words; open class nonterminals are mapped to a set of words or phrases extracted from story text. Mental states are from the mental states model mentioned in the text. Only part of the grammar is shown for illustrative purposes.*

Underlying process: Select a character [not shown] and construct question about something that character did or experienced in the story
Say: *Ask a question about the town mouse*
Listen for: [the grammar in Figure 2]

Figure 3: *Dialogue strategy for generating questions.*

## 3. Design of spoken dialogue to teach morphology

As part of current research on automatic tutoring for children's reading, we 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. These dialogues had several characteristics:

A. The time frame for intervention scenarios is relatively short, compared to scenarios for core words such as *slender* and *anxious*, where more extensive vocabulary exercises are warranted.

B. The goal of the vocabulary exercise is not core vocabulary development per se, but rather some related learning goal; in this case current text comprehension and/or familiarization with morphology in preparation for later learning.

C. The words are likely to be somewhat rare and have few distinct senses (*subvocal*, rather than *set*).

The idea, therefore, was to develop a spoken dialogue based intervention that provides relatively short interventions on words, 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 4 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 4: *Dialogue strategy for learning morphology*

**Dialogue evaluation.** These dialogue strategies for questions and for morphology have been through preliminary field tests, and evaluation of the results is in preparation, but evaluation of these specific strategies is not the focus of this paper.

## 4. Design principle: Elicit educational, predictable responses by teaching a parameterized process

The key to both the comprehension strategy dialogue and the morphology dialogue is to design the dialogue to elicit educational, predictable responses from the student. This can be generalized as follows: *Design a concise, easily learned parameterized process whereby a series of dialogue system prompts and a student mental task result in student production of a predictable spoken response.* The specific process will vary from task to task; the key is that the process require work on the part of the student and that the process yields automatically recognizable spoken output.

In the case of the comprehension strategy dialogue, the idea is to teach a process **PQ** for generating a question:

| **PQ(Character, Wh-word, Action)** | Example |
|---|---|
| **select(Character)** | Character=*the mouse* |
| **select(Wh-word)** | Wh-word=*why* |
| **select(Action)** | Action=*freeze* |
| **say: *wh- did character action?*** | *Why did the mouse freeze?* |

Multiple correct responses are possible, yet the outcome set is constrained enough to be predictable.

In the case of the vocabulary dialogue, the process **PV** is:

**PV(Word, Prefix, Stem, PrefixMeaning)** Example
**hear(Word)** Word=*reinvent*
**hear(Prefix *here means* PrefixMeaning)** Prefix=*re-*
PrefixMeaning=*again*
**say: PrefixMeaning Stem** *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 could rely on the student to supply the prefix meaning, making the student do more work when appropriate.

Several technical challenges present themselves; while the focus of this paper is not on solving any specific technical challenge for one particular dialogue, it is worth enumerating them here. The first is to be able to predict the correct outcome(s), for example by using a small (perhaps finite) language model. The second is to be able to match the predicted outcome, for example by testing whether the predicted outcome matches the output of a less-constrained language model. The third is to be able to identify incorrect responses such as silence, noise, background speech, or on-topic but wrong utterances. We acknowledge that the detection of incorrect responses (modeled perhaps with malrules) is a key part of the system design process, yet it is outside the scope of this paper, as is the design of the desired system responses to learner input.

## 5. Application of the design principle to other aspects of reading tutoring

How well does this design principle apply to other areas of reading tutoring? For **capturing narrations**, the process is producing a reading of the words in the sentence, in order, without gaps; the prompt is simply a spoken prompt to narrate the sentence; the desired response is to read the sentence fluently aloud. Here we rely on the reader's knowledge of the story (which he or she wrote or typed in) in order to make the task feasible. For **fluency**, the process is producing a fluent oral reading of the sentence, and the desired response is a fluent reading (by comparison with adult narration, for example.) For **word identification**, the process is to pronounce a word correctly in response to the printed word.

There are also negative examples: reading dialogues whose design did not take this principle into account (since the principle had not been identified yet). For **comprehension**, free-form responses with known target key word(s) aren't predictable enough for wordspotting. For **fluency**, reading a paragraph is not predictable enough to track imperfect reading.

## 6. Application of the design principle to reanalyze language learning dialogues

How well does this design principle describe language-related tasks other than reading? We discuss concept learning, grammar exercises and interaction exercises.

**Concept learning.** One interesting and timely example of language learning is aviation English [e.g. 11; current commercial offerings include Carnegie Speech, Berlitz, and DynEd]. Aviation English is a subset of English that is defined for use in international air travel; much of the work in this area appears to be commercial rather than academic due perhaps to the very large market potential. Certain aspects of existing aviation English training courses can be captured by our design principle. For example, DynEd's exercises "Commands and Questions" (http://www.dyned.com/products/ae/) present a combination of textual and visual prompts and then listen for student responses (Figure 5). Here the underlying process is to produce the name of the appropriate instrument(s) in response to a combination of textual and visual cues.


Figure 5: *DynEd Commands and Questions. The combination of a textual and visual stimulus in the prompt may serve to enhance response predictability.*

Besides concepts and vocabulary, **grammar** exercises have also been the target of spoken dialogues (as in the offering from Carnegie Speech for aviation English.) A key early example is the Subarashii system [12] which included constrained exercises with a hidden list of correct choices: "The goal of a constrained exercise is to train the student for encounters. Constrained exercises may address sentence negation, the numbering system, telling time, identifying colors, and so forth. In the course of constrained exercises, the student is likely to make a number of errors, grammatical and otherwise, that are commonly made by English learners of Japanese. The computer can track these errors and, wherever possible, provide the student with an explanation of errors they make". Here the underlying process is to produce one of a small set of responses that show mastery of a certain grammatical construct – if the student produces the utterance correctly that is taken as evidence that the grammatical structure has been learned, and grammatical mistakes are likewise evidence of a need for assistance. The prompt includes the overall context of the dialogue (the "encounter"), spoken Japanese, written English, and a picture of the virtual interlocutor.

**Conversation** exercises have also been used in language tutoring for training of culturally appropriate interaction. Early speech-based systems used on-screen menus to constrain choices [13]. One of the premier examples of this kind of training, the Tactical Language and Culture Training System, includes in each course "a set of interactive Skill Builder lessons, focusing on particular communicative skills" [14]. The utterance formation exercises (written prompt, spoken response) in particular appear to be well described by our design principle. For example, in Johnson and Valente's Figure 1, where the user is expected to generate a culturally appropriate farewell, the prompt is an image of a Chadian woman, plus the text: "It's late morning. You just met Halmé and now it's time to say goodbye. Say something appropriate to her." The tutoring system listens for correct and incorrect responses described by "a grammar-based model built from phrases extracted from the authored content" which includes "alternative ways of phrasing these utterances, some of which are well-formed and some of which may illustrate common learner errors" [JV 2008 p. 5], including (e.g.) *salut.* Here the underlying process is to produce a culturally appropriate response given strong visual and textual cues.

## 7. Application of the design principle to tasks other than language tutoring

How well does this design principle describe tutoring tasks other than language learning? We give two hypothetical examples of how our design principles could be expanded to non-language tasks in well-studied tutoring domains: the FOIL mnemonic in algebra, and applying rules in physics.

**The FOIL mnemonic in algebra** is a memory device to help students remember how to multiply two terms together: First, Outer, Inner, Last, as in $(3x+2)(2x+1) = 6x^2+3x+4x+2$. Using our dialogue design principle, we can directly construct the following dialogue for teaching FOIL (Figure 6).

> Underlying process: Apply FOIL rule to algebraic term multiplication to produce an unreduced form of the result.
> Show: **(3x+2)(2x+1)**
> Say: *Multiply these terms with FOIL - First, Inner, Outer, Last - and say each part in order.*
> Listen for:
> (first is) six x squared
> (outer is | plus) three x
> (inner is | plus) four x
> (last is | plus) two

Figure 6: *Proposed dialogue strategy for FOIL Rule.*

**Applying rules in physics** in order to calculate quantities is an integral part of solving physics problems. Using our dialogue design principle, we immediately construct the following dialogue for figuring out acceleration using the famous formula F=ma (Figure 7).

> Show: **[A diagram with a mass labeled 3 kg and force labeled 6 N (3 kg m/s$^2$).]**
> Say: *Apply F=ma to figure out the acceleration by completing this sentence: The acceleration equals* – [end with rising intonation]
> Listen for *two meters per second squared* or:
> (The acceleration equals)
> three newtons divided by six kilograms

Figure 7: *Proposed dialogue strategy for F=ma.*

## 8. Discussion and Conclusion

In this paper, we have laid out a design principle for applying spoken dialogue to intelligent tutoring: Elicit predictable response by first teaching a parameterized template or process for mapping prompt to response; a student who learns that process with predictable outcome(s) will produce the response when given the prompt. We have described its basis in reading tutoring in terms of reading aloud, the comprehension strategy of self-generated questions, and the vocabulary strategy of morphology. We have also discussed how this design principle explains other computer-assisted language learning methods with demonstrated effectiveness in the areas of concept learning, grammar, and conversation. Finally, we have illustrated how this principle can be applied to generate novel dialogues in two important non-language domains: algebra and physics. In a sense, this design principle aims to teach people how to produce speech that demonstrates mastery of the material, rather than having the system try to learn how to recognize arbitrarily difficult speech; in this respect it is similar to the Universal Speech Interface project [15]. While not all aspects of tutoring will be covered by such dialogues, we contend that this design principle describes a variety of interactions for which existing examples have had demonstrated effectiveness in computer-assisted language learning, and also has potential for widespread, effective use in other domains as well.

## 9. Acknowledgements

We would like to thank the Institute of Education Sciences (IES) for funding Project LISTEN.

## 10. References

[1] Russell, M., and D'Arcy, S., "Challenges for computer recognition of children's speech", *Proceedings of SLaTE-2007*, 108-111.

[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] 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.

[4] 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.

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

[6] 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.

[7] Neri, A., Cucchiarini, C., Strik, H. "ASR-based corrective feedback on pronunciation: does it really work?" Proceedings of Interspeech 2006.

[8] 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.

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

[10] NRP. "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." Washington, DC. 2000. http://www.nichd.nih.gov/publications/nrppubskey.cfm

[11] Dourmap, L, and Truillet, P. "Vocal interaction and air traffic management: The VOICE project", HCI-AERO 2004.

[12] Bernstein, J., Najmi, A., and Ehsani, F. "Subarashii: Encounters in Spoken Japanese Language Education", CALICO Journal 16(3):361-384, 1999.

[13] Holland, V. M., Kaplan, J.D., and Sabol, M.A., "Preliminary tests of language learning in a speech-interactive graphics microworld", *CALICO Journal*, 16(3):339-359.

[14] Johnson, W. L. and Valente, A., "Tactical Language and Culture Training Systems: Using Artificial Intelligence to Teach Foreign Languages and Cultures", Proceedings of IAAI 2008.

[15] Rosenfeld, R., Olsen, D., and Rudnicky, A. Universal Speech Interfaces. *Interactions*, VIII(6), 2001, pp. 34-44.

Carnegie Speech's speech recognition-based grammar exercises mentioned in descriptive material for "Climb Level 4": http://www.maycoll.co.uk/pdfs/climb-level4.pdf