Positive and negative verbal feedback for Intelligent Tutoring Systems

Barbara Di Eugenio [a,1] Xin Lu [a] Trina C. Kershaw [a] Andrew Corrigan-Halpern [a] Stellan Ohlsson [a]

[a] *University of Illinois, Chicago, USA*

Abstract. We built three different versions of an ITS on a letter pattern extrapolation task: in one version, students only receive color-coded feedback; in the second, they receive verbal feedback messages when they perform correct actions, and in the third, when they make a mistake. We found that time on task and number of errors are predictive of performance on the post-test rather than the type of feedback.

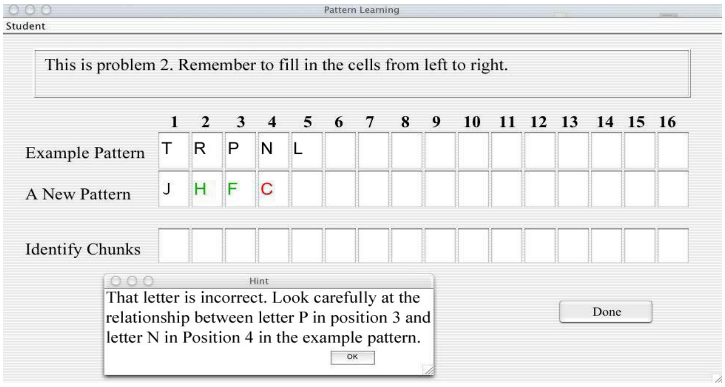
Keywords. Intelligent Tutoring Systems. Natural Language feedback.

1. Introduction and motivation

Research on the next generation of Intelligent Tutoring Systems (ITSs) [2,3,4] explores Natural Language (NL) as one of the keys to bridge the gap between current ITSs and human tutors. In this paper, we describe an experiment that explores the effect of simple verbal feedback that students receive either when they perform a correct step or when they make a mistake. We built three different versions of an ITS that tutors students on extrapolating a complex letter pattern [7], such as inferring MEFMGHM from MABMCDM. In the *neutral* version of the ITS the only feedback students receive is via color coding, green for correct, red for incorrect; in the *positive* version, they receive feedback via the same color coding, and verbal feedback on correct responses only; in the *negative* version, they receive feedback via the same color coding, and verbal feedback on incorrect responses only. In a between-subject experiment we found that, even if students in the verbal conditions do perform slightly better and make fewer mistakes, these differences are not significant. Rather, it is time on task and number of errors that are predictive of performance on the post-test.

This work is motivated by two lines of theoretical inquiry, one on the role of feedback in learning [1], the other, on what distinguishes expert from novice tutors [8]. In another experiment in the letter pattern domain, subjects were individually tutored by three different tutors, one of which had years of experience as a professional tutor. Subjects who were tutored by the expert tutor did significantly better on one of the two problems in the post-test, the more complex one. The content of the verbal messages in our ITSs is based on a preliminary analysis of the language used by the expert tutor.

[1] Correspondence to: B. Di Eugenio, Computer Science (M/C 152), University of Illinois, 851 S. Morgan St., Chicago, IL, 60607, USA. Email: bdieugen@cs.uic.edu.



This is problem 2. Remember to fill in the cells from left to right.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Example Pattern	T	R	P	N	L											
A New Pattern	J	H	F	C												
Identify Chunks																

That letter is incorrect. Look carefully at the relationship between letter P in position 3 and letter N in Position 4 in the example pattern.

Figure 1. The *negative* ITS, that provides verbal feedback on mistakes

2. Method and Results

Our three ITSs are model-tracing tutors, built by means of the Tutoring Development Kit [6]. Fig. 1 shows the interface common to all three ITSs. The *Example Pattern* row presents the pattern that needs to be extrapolated; the *A New Pattern* row is used to enter the answer – the first cell of this row is filled automatically with the letter the extrapolation must start from; the *Identify Chunks* row can be used to identify chunks, as a way of parsing the pattern. If seen in color, Fig. 1 also shows that when the subject inputs a correct letter, it turns green (H, F), and when the subject makes a mistake, the letter turns red (C).

We ran a between-subjects study in which each group of subjects (positive [N = 33], negative [N = 36], and neutral [N = 37]) interacts with one version of the system. All subjects first received instructions about how to interact with the ITS. The positive and negative groups were not informed of the feedback messages they would receive. All subjects trained on the same 13, progressively more difficult, problems, and then received the same post-test consisting of 2 patterns, each 15 letters long. Subjects see the same pattern for 10 trials, but must continue the pattern starting with a different letter each time. Post-test performance is the total number of letters that subjects enter correctly across the 20 trials (a perfect score is 300).

	Post-test score	Time	Errors
Positive	154.06	42.68	18.91
Negative	141.83	45.52	14.69
Neutral	134.62	42.02	21.89

Table 1. Means for the three groups

Means for each condition on post-test scores, time spent in training, and number of errors are shown in Table 1. Subjects in the two verbal conditions did slightly better on the post-test than subjects that did not receive any verbal feedback, and they made fewer mistakes. Further, subjects in the positive condition did slightly better than subjects in the negative condition on the post-test, although subjects in the negative condition made fewer mistakes. However, none of these differences is significant.

A linear regression analysis was performed with post-test scores as the dependent variable and condition, time spent in training, and number of errors as the predictors. The overall model was significant, $R^2 = .16$, $F(3, 102) = 6.52$, $p < .05$. Time spent in training ($\beta = -.24$, $t(104) = -2.51$, $p < .05$) and number of errors ($\beta = -.24$, $t(104) = -2.53$, $p < .05$) were significant predictors, but condition was not a significant predictor ($\beta = -.12$, $t(104) = -2.53$, $p > .05$).

Hence, we can explain variation in the post-test scores via individual factors rather than by feedback condition. The more time spent on training and the higher number of errors, the worse the performance. However, it would be premature to conclude that verbal feedback does not help, since there may be various reasons why it was not effective in our case. First, students may have not really read the feedback, especially in the positive condition in which it may sound repetitive after some training [5]. Second, the feedback may not be sophisticated enough. In the project DIAG-NLP [2] we compared three different versions of an ITS that teaches troubleshooting skills, and found that the version that produces the best language significantly improves learning. The next step in the letter pattern project is indeed to produce more sophisticated language, that will be based on a formal analysis of the dialogues by the expert tutor. On the other hand, it may well be the case that individual differences among subjects are more predictive of performance on this task than type of feedback. We will therefore also explore how to link the student model with the feedback generation module.

Acknowledgments. This work is supported by awards CRB S02 and CRB S03 from UIC, and by grant N00014-00-1-0640 from the Office of Naval Research.

References

[1] A. Corrigan-Halpern and S. Ohlsson. Feedback effects in the acquisition of a hierarchical skill. In *Proceedings of the 24th Annual Conference of the Cognitive Science Society*, 2002.

[2] B. Di Eugenio, D. Fossati, D. Yu, S. Haller, and M. Glass. Natural language generation for intelligent tutoring systems: a case study. In *AIED 2005, the 12th International Conference on Artificial Intelligence in Education*, 2005.

[3] M. W. Evens, J. Spitkovsky, P. Boyle, J. A. Michael, and A. A. Rovick. Synthesizing tutorial dialogues. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society*, pages 137–140, 1993.

[4] A.C. Graesser, N. Person, Z. Lu, M.G. Jeon, and B. McDaniel. Learning while holding a conversation with a computer. In L. PytlikZillig, M. Bodvarsson, and R. Brunin, editors, *Technology-based education: Bringing researchers and practitioners together*. Information Age Publishing, 2005.

[5] Trude Heift. Error-specific and individualized feedback in a web-based language tutoring system: Do they read it? *ReCALL Journal*, 13(2):129–142, 2001.

[6] Kenneth R. Koedinger, Vincent Aleven, and Neil T. Heffernan. Toward a rapid development environment for cognitive tutors. In *12th Annual Conference on Behavior Representation in Modeling and Simulation*, 2003.

[7] K. Kotovsky and H. Simon. Empirical tests of a theory of human acquisition of information-processing analysis. *British Journal of Psychology*, 61:243–257, 1973.

[8] S. Ohlsson, B. Di Eugenio, A. Corrigan-Halpern, X. Lu, and M. Glass. Explanatory content and multi-turn dialogues in tutoring. In *25th Annual Conference of the Cognitive Science Society*, 2003.