Book Title
Book Editors
IOS Press, 2003

Assessing Learning in a Peer-Driven Tutoring System

Ari Bader-Natal [1], Jordan Pollack
DEMO Lab, Brandeis University

Abstract.
In many intelligent tutoring systems, a detailed model of the task domain is constructed and used to provide students with assistance and direction. Reciprocal tutoring systems, however, can be constructed without needing to codify a full-blown model for each new domain. This provides various advantages: these systems can be developed rapidly and can be applied to complex domains for which detailed models are not yet known. In systems built on the reciprocal tutoring model, detailed validation is needed to ensure that learning indeed occurs. Here, we provide such validation for SpellBEE, a reciprocal tutoring system for the complex task domain of American-English spelling. Using a granular definition of response accuracy, we present a statistical study designed to assess and characterize student learning from collected data. We find that students using this reciprocal tutoring system exhibit learning at the word, syllable, and grapheme levels of task granularity.

American-English spelling is a domain known to be difficult to fully codify [5]. It serves as the task domain addressed by the SpellBEE reciprocal tutoring system, which is the first of a growing suite built on the BEEweb model [2]. SpellBEE has been publicly available online (at SpellBEE.org) for the past three years and has, to date, collected data from over 17,000 active participants engaged in over 22,000 completed peer-tutoring sessions. Participants primarily consist of American students in grades 2 through 8.

Student interactions in the BEEweb model are strictly governed by a two-student reciprocal tutoring arrangement, similar to prior classroom-based protocols compiled by O'Donnell and King [6] and the computer-based protocols of Chan and Chou [3]. This reciprocal tutoring protocol determines the structured flow of interactions between a pair of students. A game theoretic framework, the "Teacher's Dilemma," serves as a motivational mechanism within the game. Each step in the reciprocal tutoring protocol corresponds with some aspect of this game. Details on the BEEweb model and the Teacher's Dilemma can be found in our earlier work [1,2], so the protocol is only briefly described here: Once a pair of students has elected to engage in a collaboration session, they proceed through a sequence of steps (repeated once during each turn of the game), alternating between playing the roles of "tutor" and "tutee." At the first step, a student is asked to construct a challenge to pose to their partner (In SpellBEE, this construction task takes the form of choosing from a short list of words randomly selected from a 3000+ word dictionary.) In the second step, the student attempts to solve the challenge that was posed by their partner in the previous step (In SpellBEE, the student sees a sentence context

[1]Correspondence to: Ari Bader-Natal, Brandeis University, Computer Science Department MS018. Waltham, MA 02454. USA. Tel.: +1 781 736 3366; Fax: +1 781 736 2741; E-mail: ari@cs.brandeis.edu.

for the word with the challenge word hidden, hears the entire sentence read aloud by text-to-speech software, and then types a response spelling into a text field.) When the student submits this response, they receive feedback on its accuracy and are shown the correct response if they made a mistake. Finally, the student is presented with feedback on how their partner fared on the problem that they selected in the first step, which can be useful when constructing future challenges.

Given the peer-driven nature of this design, we need to validate that students using SpellBEE are indeed learning and improving at spelling. To do this, we first define a fine-grained concept of response accuracy and select two domain-appropriate notions of sub-problem structure to examine, *syllables* and *graphemes*.[1] We then run statistical tests to identify if and how SpellBEE students improve at spelling, both overall and with respect to these two sub-word structures.

Past analysis of SpellBEE has been based on a scalar measure of challenge difficulty and a dichotomous measure of response accuracy [2]. We refer to this accuracy measure here as *whole-word accuracy*, which is defined in terms of a student's response (spelling), r, to the (word) challenge posed to them, c:

$$\mathcal{A}(c,r) = \begin{cases} 0 & \text{if response } r \text{ is } \textit{not} \text{ a correct solution to challenge } c \\ 1 & \text{if response } r \text{ is a correct solution to challenge } c \end{cases}$$

In order to gain a deeper understanding of the nature of student improvements within the spelling domain, we need to analyze spelling accuracy at a finer grain. Hanna et al. [5] thoroughly detailed the role and regularity of phoneme-grapheme correspondences in American-English spelling, and we draw upon in this work by examining spelling accuracy at the levels of graphemes and syllables. We introduce a concept of *sub-word accuracy*, defined with respect to a sub-word structure s, such as a grapheme or syllable:

$$\mathcal{A}'(c,r,s) = \begin{cases} 0 & \text{if sub-problem } s \text{ of challenge } c \text{ was } \textit{not} \text{ correctly solved in } r \\ 1 & \text{if sub-problem } s \text{ of challenge } c \text{ was correctly solved in } r \end{cases}$$

These two definitions of accuracy can be leveraged to construct statistical tests for learning. As SpellBEE does not currently incorporate pre- and post-testing, we rely on McNemar's test to examine the effect of SpellBEE usage on spelling improvement. This is a non-parametric statistical method that tests for change in a dichotomous trait within a group of subjects before and after an intervention [4]. When a student sees some challenge c at time t_i and later at t_j, the change from $\mathcal{A}(c,r_i)$ to $\mathcal{A}(c,r_j)$ can indicate learning. For some fixed challenge c, let Δ be the number of students for whom $\mathcal{A}(c,r_i) < \mathcal{A}(c,r_j)$, and let ∇ be the number of students for whom $\mathcal{A}(c,r_i) > \mathcal{A}(c,r_j)$.[2] McNemar's test uses Δ and ∇ to test the association between SpellBEE usage and response accuracy[3] (using the statistic: $\chi^2_{McNemar} = \frac{|\Delta-\nabla|^2}{\Delta+\nabla}$.) With this approach, we find that the association between whole-word spelling accuracy and SpellBEE usage is significant, with accuracy improving with usage (Yates' continuity-corrected $\chi^2 = 28.2031$, $df = 1$, $p < 0.001$, odds ratio $= 1.9891$).

Based on the finer-grained $\mathcal{A}'$ accuracy, we can use McNemar's test to see if students are learning the spelling of sub-word structures that occur in many different words.

[1] A *phoneme* is the smallest unit of sound in a language, and a *grapheme* is the written form of the phoneme.
[2] If more than one comparison is available for a student, the one with the most elapsed time ($t_j - t_i$) is used.
[3] Note that the number of students for whom $\mathcal{A}(c,r_i) = \mathcal{A}(c,r_j)$ is not used.

Table 1. Graphemes are classified according to how student spelling accuracy changed during SpellBEE usage.

Improved ($p < 0.05$)	No Significant Change ($p \geq 0.05$)	Worsened ($p < 0.05$)
A, C, CC, CE, CQ, CQU, CT, D, DG, E, ED, EI-E, EIGH, EN, ES, F, G, GH, GI, H, I, I-E, IA, IA-E, IGH, IN, J, K, KN, L, M, N, NG, O, OL, ON, OO, OW, OW-E, P, PP, PT, Q, QU, R, S, SI, SSI, ST, T, TH, TI, U, U-E, W, WH, X, Y	A-E, AI, AI-E, AL, AU, AW, AY, B, CH, CI, CK, DD, DI, E-E, EA, EA-E, EE, EE-E, EI, EL, EO, ET, EW, EY, –EY, FF, FT, GG, GN, GU, GUE, IE, IE-E, IL, LD, LE, LL, LV, MB, MM, MN, NN, O-E, OA, OI, OU, OUGH, OWE, RR, SC, SCI, SH, SL, SS, SW, TCH, TT, UE, UI, UI-E, V, WR, Z	*none*

For words c_x and c_y containing a common *grapheme* s, we now let Δ count the students for whom $\mathcal{A}'(c_x, r_i, s) < \mathcal{A}'(c_y, r_j, s)$ and let ∇ count the students for whom $\mathcal{A}'(c_x, r_i, s) > \mathcal{A}'(c_y, r_j, s)$. Table 1 shows that at the $\alpha = 0.05$ level, we found 58 graphemes for which student spelling accuracy significantly changed for the better, 63 graphemes for which no significant change was observed, 0 graphemes for which student spelling accuracy significantly changed for the worse, and 50 graphemes for which not enough data was available to use the test (i.e. $\Delta + \nabla < 10$). Similarly, when we examine changes in response accuracy over time at the granularity of *syllables* as sub-structure s (for which enough data was available to use McNemar's method), we find that students significantly improved on 79 syllables, exhibited no significant change on 304 syllables, and significantly worsened on 0 syllables.

Notably, we observed no syllable or grapheme for which student spelling significantly worsened after SpellBEE usage, and many for which student spelling significantly improved. Results from all three levels of analysis granularity support the conclusion that students are improving at the spelling task with usage of the tutoring system, and the analysis allows us see how this progress is distributed across sub-problem structures. Finally, we find that sub-word accuracy can be used as the basis for a principled statistical validation of learning in the SpellBEE reciprocal tutoring system. By choosing domain-appropriate definitions of sub-problem accuracy, the methodology used here can be applied to analyzing student learning other tutoring systems.

References

[1] Ari Bader-Natal and Jordan Pollack. Motivating appropriate challenges in a reciprocal tutoring system. In C.-K. Looi, G. McCalla, B. Bredeweg, and J. Breuker, editors, *Proc. of the 12th Intl. Conf. on AI in Education (AIED-2005)*, pages 49–56, Amsterdam, July 2005. IOS Press.

[2] Ari Bader-Natal and Jordan Pollack. BEEweb: A multi-domain platform for reciprocal peer-driven tutoring systems. In M. Ikeda, K. Ashley, and T.-W. Chan, editors, *Proc. of the 8th Intl. Conf. on Intelligent Tutoring Systems (ITS-2006)*, pages 698–700. Springer, June 2006.

[3] Tak-Wai Chan and Chih-Yueh Chou. Exploring the design of computer supports for reciprocal tutoring. *International Journal of Artificial Intelligence in Education*, 8:1–29, 1997.

[4] B. S. Everitt. *The Analysis of Contingency Tables*. Chapman and Hall, 1977.

[5] Paul Hanna, Jean Hanna, Richard Hodges, and Edwin Rudorf. Phoneme-grapheme correspondences as cues to spelling improvement. Report OE-32008, Office of Education, 1966.

[6] Angela M. O'Donnell and Alison King, editors. *Cognitive Perspectives on Peer Learning*. Lawrence Erlbaum Associates, 1999.