Putting a/the stake in the ground: Making a priori predictions of student learning

Ruth Wylie[1] , Kenneth Koedinger[1], and Teruko Mitamura[2]

[1] Human-Computer Interaction Institute, [2] Language Technologies Institute
School of Computer Science, Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, Pennsylvania 15213
{rwylie, koedinger, teruko}@cs.cmu.edu

Abstract. A fundamental challenge when designing interfaces for instructional systems is determining how much scaffolding or assistance they should provide. Less assistance may encourage deeper processing in the best case but yield unproductive floundering or failure in the worst case. Working within the domain on the English article system (e.g. *the* dog vs. *a* dog), we provide an approach for making a priori predictions about learning gains. The theoretical descriptions of learning event spaces combined with empirical data from a small think-aloud study (n=6) predict that the process of error detection is extraneous; thus, the higher-assistance menu interface, which scaffolds the error detection process, should produce greater learning gains than the lower-assistance edit interface which requires students to both detect and correct errors.

Keywords: Instructional design, Assistance Dilemma, CALL, English article system, Learning Event Space

1 Introduction

A fundamental challenge in instructional design is creating tasks that are at an appropriate difficulty level. Vygotsky's Zone of Proximal Development, Krashen's Input Hypothesis, and Cognitive Load Theory all suggest that students learn most when problems are neither too challenging nor too easy. While it's hard to argue against this claim, the difficulty arises when one tries to apply the theories. For example, in the absence of data from learning studies, it is unclear how to determine a student's "actual developmental level" [1], which input is just beyond a learner's current level of comprehension [2], or whether a cognitive process is extraneous or germane [3]. While randomized controlled classroom-based learning studies remain the gold standard, they are not always practical or even possible to conduct. Thus, this paper presents an approach that combines cognitive theory with a small set of performance data to make a priori predictions about learning gains and inform instructional design. We reframe the question in terms of the assistance dilemma: "How should learning environments balance information or assistance giving and withholding to achieve optimal student learning?" [4], and construct learning event spaces for two interfaces designed to teach the English article (a, an, the, null) system. Finally, we use think-aloud protocols from six English language learners (ELLs) to infer how students make English article decisions and inform our predictions.

1.1 Domain: English articles

The English article system is one of the most difficult aspects of grammar to teach second language learners [5]. The rules are complex and there are many exceptions. However, there is evidence that explicit, systematic teaching of article rules results in higher learning gains compared to the control [6] and at least one intelligent tutoring system has been built to explicitly teach students the rules [7].

A common critique of this domain is that successful article use is often not required for communication to occur (e.g. Listeners usually understand the phrase "Please pass me pencil" even though "Please pass me the pencil" is correct.) However, correct article usage becomes much more important in writing, where readers cannot rely on extralinguistic cues (e.g. pointing) for clarification. In addition, article errors fall into the category of nonnative-like errors and can affect the overall credibility of the work [8].

For this study, we've chosen to focus on editing rather than an open-ended task such as essay writing. While writing is an important part of second language acquisition, editing is also an important skill in order for students to be able to successfully use articles in their own writing. Further, our research goal is to understand which tasks produce greater learning gains, and the large variety in responses from open-ended tasks makes it difficult, if not impossible, to perform rigorous analyses.

1.2 High and Low Assistance Interfaces

We built two interfaces, with varying degrees of assistance, using the Cognitive Tutoring Authoring Tools (CTAT) [9]. The first interface, the menu interface, is similar to cloze, or fill-in-the-blank, activities found in many English as a Second Language (ESL) textbooks. It provides higher assistance because students are not required to identify where errors exist; they simply choose a response for each box.

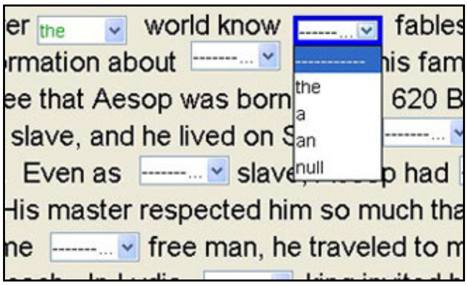
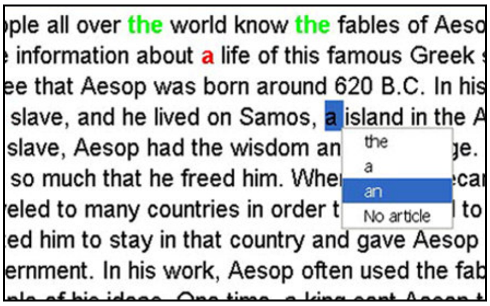


Fig. 1: Using the higher assistance, menu interface, students select the appropriate article from each menu.

The second interface, the editing interface, provides lower assistance because students must both detect the error and produce the correct response. Students can make changes by inserting, removing, or replacing articles anywhere in the text. However, only articles can be edited thus preventing students from completing rewriting sentences in order to avoid unknown grammar constructions (Figure 2).



ple all over the world know the fables of Aeso
information about a life of this famous Greek
ee that Aesop was born around 620 B.C. In his
slave, and he lived on Samos, a island in the A
slave, Aesop had the wisdom an ge.
so much that he freed him. Whe cal
eled to many countries in order t l to
ed him to stay in that country and gave Aesop
ernment. In his work, Aesop often used the fab

Fig. 2: Using the lower assistance editing interface, students find *and* correct article errors.

The assistance dilemma question is thus, is the process of error detection extraneous and therefore should be performed by the interface (as in the menu condition) or is it a germane process and should be performed by the student (as in the edit condition)? In answering this question, we will next consider the learning event space for each of the two interfaces.

2 Learning Event Spaces

A first step in resolving the assistance dilemma is to understand the paths students may follow when completing the task. Learning Event Spaces communicate all possible ways of completing the task and include optimal routes (e.g. those that encourage deep processing of the target skills) as well as suboptimal routes (e.g. a student guesses and picks the right answer by chance). This information can be useful for researchers in order to explain learning effects (e.g. Students who followed Path A learned more than students who followed Path B) as well as for instructional designers who can use path effects to design systems that afford the following of certain paths over others [10]. It's important to note that the following learning event spaces are for the specific tasks (the menu interface and the edit interface) and not for the domain in general. For example, since our tasks do not require students to make decisions for noun phrases that are exceptions to rules or idiomatic expressions, the learning event spaces do not include paths for making these decisions.

2.1 Learning Event Space: Higher-Assistance Interface

When students complete the task using the higher-assistance menu interface, they read the paragraph until they reach a menu and then select the appropriate article by retrieving and applying the relevant rule, choosing the article that “sounds right”, or guessing (Table 1).

Table 1: Learning Event Space for Higher-Assistance Menu Interface

- H-1. Student reads text and reaches a menu
- H-2. Student makes an article selection by:
  - 2.1 retrieving and applying the relevant rule
  - 2.2 choosing the article that “sounds right
  - 2.3 guessing

2.2 Learning Event Space: Lower-Assistance Interface

The learning event space for the lower-assistance edit interface is naturally more complex than the learning event space for the menu interface given the additional task of error detection. (Table 2). In order to make an edit, students read the text until they reach: a phrase that sounds incorrect (step 1.1), an article (a, an, or the) (step 1.2), or a noun (step 1.3). While all articles will eventually be followed by a noun including both options 1.2 and 1.3 may seem redundant. However, they are listed as separate paths because the presence of an article may serve as a salient reminder, prompting students to evaluate noun phrases that contain articles more often than noun phrases that initially contain no article. In the case of paths 1.2 and 1.3, students next evaluate the noun phrase by either retrieving and applying a relevant rule or by guessing that it is wrong. Finally, students make their article selection by retrieving and applying the relevant rule, choosing the article that “sounds right”, or by guessing. Note the article selection process for the lower-assistance interface (step L-3) is identical to the selection process for the higher-assistance interface (step H-2).

Table 2: Learning Event Space for Edit Interface (Lower Assistance)

- L-1 Student reads the text until and reaches:
  - 1.1 a phrase that “sounds wrong”, go to step E-3
  - 1.2 an article
  - 1.3 a noun phrase
- L-2 Student evaluates whether what is present is correct by:
  - 2.1 retrieving and applying the relevant rule
  - 2.2 guess that it is wrong
- L-3 Student makes an article selection by:
  - 3.1 retrieving and applying the relevant rule
  - 3.2 choosing the article that “sounds right”
  - 3.3 guessing

2.3 Implications for Learning

Determining whether the process of error detection is extraneous or germane depends largely on the paths students follow as they use the lower-assistance edit interface. If students are systematically evaluating each noun phrase (NP) to detect errors (step L-2), then the lower-assistance edit interface provides more opportunities to retrieve and apply rules than the higher-assistance menu interface does. Specifically, since students in the lower-assistance condition are responsible for detecting errors, they need to evaluate all noun phrases in the paragraph – both those that contain an error (initially incorrect) and those that do not (initially correct). However, since students using the higher-assistance menu interface are only required to make selections where a menu occurs and since menus only occur for approximately 20-30% of all noun phrases, students using the higher-assistance menu interface have fewer practice opportunities with the rules. Thus, if students are evaluating each noun phrase when using the edit interface (L-2), this additional practice should result in greater learning as compared to students who complete the same paragraph using the menu interface.

However, if students are detecting errors using an implicit "sounds wrong" strategy (step L-1.1), they would be skipping the evaluation step (L-2) and thus not receiving additional practice with retrieving and applying the rules, and as such greatly reducing the potential benefits of the edit interface. In fact, if students are relying on a "sounds wrong" strategy, they are likely able only to identify the errors for which they already have an implicit understanding of what the correct response should be. That is, the lower-assistance edit interface might promote students practicing only the rules they already know. On the other hand, students using the higher-assistance menu interface are forced to make a selection for all noun phrases with a menu, even those with which they may be unfamiliar. Thus, if students are following this path (L-1.1), the higher-assistance menu interface should provide more practice to students and result in greater learning gains.

3 Empirical Validation

A study was conducted to gather data on the performance differences between the two task types as well as gather verbal protocols in order to understand which paths of the learning event space students follow when making article decisions.

3.1 Participants

Participants were recruited from Carnegie Mellon University's InterCultural Communication Center's Academic Culture and Communication (ACC) program. The program is a six week summer program designed for newly admitted non-native English speaking students. Students entering the program have high TOEFL scores (greater than 580) and at least an intermediate level of spoken fluency. ACC students are a highly educated and highly motivated group.

Study recruitment was conducted via e-mail and six students volunteered to participate (three female, three male). While not inherent in the study design, all were native Chinese speakers. The average age was 28 years old and all students had been learning English for an average of 14.4 years. Using self-report scales, participants gave an average proficiency score of 3.4 for reading, 3.0 for writing, and 2.5 for speaking (where 1 represents absolute beginner and 5 fluent).

The participants of this study represent a population that is more advanced than the overall target population of these systems. Due to the think-aloud methodology, we needed students whose English ability was high enough that they were able to verbalize what they were doing. However, because of the difficulty with this domain, we were able to use texts that were challenging for our participants.

3.2 Paragraph Content

The two problem paragraphs came from intermediate and advanced-level ESL textbooks [11] [12]. The intermediate paragraph is shorter in length and uses simpler vocabulary than the advanced paragraph does. Furthermore, the type of article use varies between paragraphs. For example, there are more instances of mass nouns in the advanced paragraph and most instances of unique-for-all usages (e.g. *the* moon, *the* sun) in the intermediate paragraph.

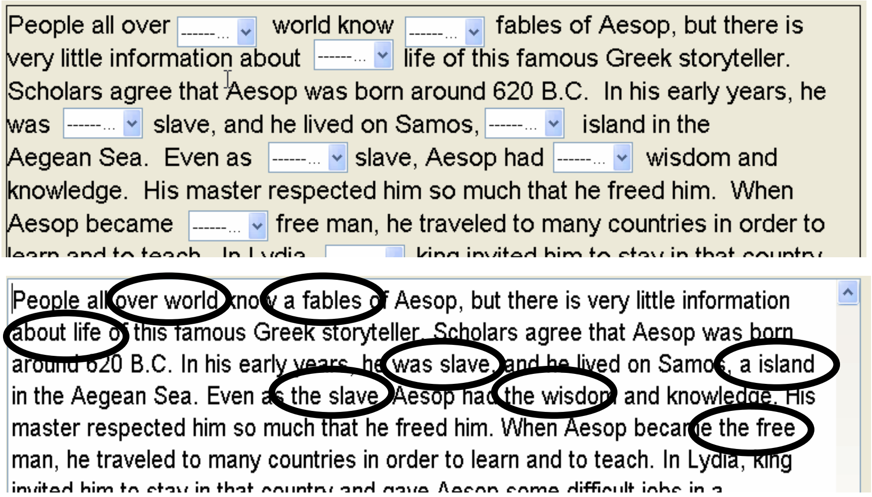
3.3 Procedure

In addressing our questions, we performed a think-aloud, lab study in which participants completed two tasks using the interfaces described above. Think-aloud methodology [13] is used to collect a verbalization of participants' thoughts while completing a task. These protocols provide great insight into the student's decision making and thought processes. While traditional think-aloud methodology prohibits the researcher from probing the participant while doing the task (e.g. following up when no explanation was given), we slightly modified the procedure and, if students made changes but did not verbalize a reason, asked students to explain why. The reason for this methodological change is because one of the goals of this study was to find out how students were making these decisions. However, it should be noted that students did not have to state an explicit rule as an explanation; implicit reasons such as "it sounds better" were acceptable and not further probed.

Students were randomly assigned to one of two groups: those in Group 1 (n=3) completed the intermediate level problem using the editing (lower assistance) interface and the advanced level problem with the menu (higher assistance) interface while students in Group 2 (n=3) did the opposite. Students were told beforehand that the paragraphs contained only article errors.

3.4 Interface Alignment

While the interfaces between the two systems require students to perform different actions (selection only for the higher-assistance menu interface; error detection and selection for the lower-assistance editing interface), when comparing performance data between the two interfaces, we looked at only the instances where students would need to make article selections for the same noun phrases, regardless of the interface with which they were working. In other words, every time a menu was inserted into the text in the menu interface, an error occurred in the text of the editing interface (Figure 3). This allows for better comparison between the two interfaces by controlling for the type of noun phrases with which students are working.



People all over ------... world know ------... fables of Aesop, but there is very little information about ------... life of this famous Greek storyteller. Scholars agree that Aesop was born around 620 B.C. In his early years, he was ------... slave, and he lived on Samos, ------... island in the Aegean Sea. Even as ------... slave, Aesop had ------... wisdom and knowledge. His master respected him so much that he freed him. When Aesop became ------... free man, he traveled to many countries in order to learn and to teach. In Lydia, king invited him to stay in that country

People all over world know a fables of Aesop, but there is very little information about life of this famous Greek storyteller. Scholars agree that Aesop was born around 620 B.C. In his early years, he was slave, and he lived on Samos, a island in the Aegean Sea. Even as the slave, Aesop had the wisdom and knowledge. His master respected him so much that he freed him. When Aesop became the free man, he traveled to many countries in order to learn and to teach. In Lydia, king invited him to stay in that country and gave Aesop some difficult jobs in a

Fig. 3: For every menu in the menu interface, an error exists in the editing interface

4 Results

As reported in [14], there were performance differences between the two interfaces, with students performing better with the higher-assistance menu interface than with the lower-assistance edit interface. Evidence from the log data generated by the tutors when combined with the think-aloud protocols suggests that when students use the edit interface, they are not systematically evaluating each noun phrase in search of errors but are relying on implicit processes for error detection.

Overall students were better at selecting articles using the menu interface than they were at identifying and correcting article errors using the editing interface.

Table 3: Student accuracy by interface and paragraph difficulty.

	Intermediate Paragraph		*Advanced Paragraph*	
	Menu Interface	*Edit Interface*	*Menu Interface*	*Edit Interface*
P1		10/11 (90.9%)	12/20 (60.0%)	
P2	9/11 (81.8%)			12/20 (60.0%)
P3		9/11 (81.8%)	17/20 (85.0%)	
P4	9/11 (81.8%)			5/20 (25.0%)
P5		5/11 (45.5%)	10/20 (50.0%)	
P6	9/11 (81.8%)			4/20 (20.0%)
Total	**27/33 (81.8%)**	**24/33 (72.7%)**	**39/60 (65.0%)**	**21/60 (35.0%)**

Since, as mentioned above in the alignment section, this data reflects places where students were required to make changes to the same noun phrase across both conditions, there are two potential explanations for why students using the edit interface did not perform as well as students who used the menu interface, either (1) Students were detecting errors but not making the correct changes or (2) students did not detect the errors. To answer this question, we look at the accuracy of student changes using the edit interface, that is, if students attempted to make a change (i.e. they correctly identified that an error existed), were they able to make the appropriate correction?

Table 4: Number of errors identified and correctly changed.

	Intermediate Paragraph		*Advanced Paragraph*	
	Errors Identified	*Correct Changes*	*Errors Identified*	*Correct Changes*
P1	10/11 (90.9%)	10/10 (100.0%)		
P2			14/20 (60.0%)	12/14 (85.7%)
P3	9/11 (81.8%)	9/9 (100.0%)		
P4			10/20 (25.0%)	5/10 (50.0%)
P5	5/11 (45.5%)	5/5 (100%)		
P6			8/20 (20.0%)	4/8 (50.0%)
Total	**24/33 (72.7%)**	**24/24 (100.0%)**	**32/60 (53.3%)**	**21/32(65.6%)**

The data in Table 4 show that compared to their overall accuracy, students are better able to select the correct article once an error has been found (100.0% vs. 72.7% for the intermediate paragraph, 65.6% vs. 53.3% for the advanced paragraph). However, the above reflects student performance only on the cases in which the noun phrases initially contained an error. Since students using the edit interface are allowed to make changes anywhere in the text, it's possible to inadvertently introduce errors by making edits to noun phrases that were initially correct. If students were frequently deciding that noun phrases were incorrect by guessing (following path L-2.2), one would expect to find several instances of students making changes to noun phrases that were initially correct. In fact, across all students and paragraphs, there were only 7 total errors that were the result of changes being made to initially-correct noun phrases (2 in the intermediate paragraph, 5 in the advanced paragraph). Thus, students are primarily making changes where errors exist and are successfully leaving initially-correct noun phrases alone, suggesting that students are either identifying incorrect noun phrases through an implicit "sounds wrong" process (path L1) or through accurately evaluating each noun phrase (path L-2.1). The verbal protocols aid in making this distinction by revealing how often students perform noun phrase evaluation. A noun phrase was coded as being evaluated if either (1) an edit was made or (2) if the protocol contained verbal evidence of evaluation (e.g. monitoring statements – " *'made from the iron'* the iron? yeah" [P2, advanced paragraph], or explicit rules " '*… sent the gold*' because this is the second time that the article mentioned gold." [P1, intermediate paragraph]).

Table 5: Averages number of noun phrases (NP) evaluated the average # of correct decisions by paragraph and for initially-incorrect and initially-correct noun phrases. (Lower-Assistance Edit Interface Only)

	Noun Phrase Initially Correct?	*Total # of NPs in Paragraph*	*Average # of NP's evaluated*	*% of NPs evaluated*	*Average # of Correct Decisions*	*% of Correct Decisions When Evaluated*
Intermediate	Yes	37	2	5.4%	1.5	75%
	No	11	8	72.7%	8	100%
Advanced	Yes	158	9.3	5.9%	7.6	81.5%
	No	22	12.3	55.9%	7	56.9%
Total	**Yes**	**195**	**11.3**	**5.8%**	**9.1**	**80.3%**
	No	**33**	**20.3**	**61.5%**	**15**	**73.9%**

The data in Table 5 show that students evaluate initially-correct noun phrases less than 6% of the time, suggesting that students are not systematically evaluating every noun phrase in the paragraph (steps L-1.2 and L-1.3 in the learning event space) but are instead detecting errors through an implicit "sounds wrong" strategy (step L-1.1).

4 Discussion

The flexibility of the lower-assistance edit interface allows two, very distinct, methods of error detection. Students can systematically evaluate each noun phrase, thus greatly increasing the number of practice opportunities to retrieve and apply article rules as compared to the higher-assistance menu interface. Or, students can detect errors using an implicit "sounds wrong" strategy. In the best case (i.e. a student detects all the errors), using the implicit strategy will provide the same number of article selection opportunities as the menu interface does. However, for every error that students fail to detect while using the lower-assistance edit interface, they are also missing an opportunity to practice error selection.

The log data and think-aloud protocols provide evidence that students are detecting errors primarily through the use of the implicit, "sounds wrong" strategy (step L-1.1), and not doing explicit, systematic evaluation. Since time spent looking for errors is time spent not practicing the rules, these results suggest that the error detection process is extraneous and should be scaffolded by the interface. More specifically, the menu interface which focuses student efforts on article selection will lead to greater learning gains than the edit interface which requires students to both detect errors and select the correct response.

The logical next step for this work is conduct the learning study in order to determine the accuracy of the combined theoretical and performance data approach. In addition to normal pre-/post-test learning measures, we plan to collect student writing samples in order to measure transfer to authentic production.

This work provides a process to understand and evaluate the manners in which students solve problems. This information is useful for both instructional design and can aid in making predictions regarding which instruction leads to greater learning.

Acknowledgements. Special thanks to Jim Rankin, Erin Walker and the PSLC DataShop team. This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305B040063 to Carnegie Mellon University and by the Pittsburgh Science of Learning Center which is funded by the National Science Foundation award number SBE-0354420. The opinions expressed are those of the author and do not represent views of the Institute or the U.S. Department of Education.

References

1. Vygotsky, L.S. (1978). Mind and society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.
2. Krashen, S.D. (1985). *The Input Hypothesis: Issues and implications.* London: Longman.
3. Sweller, J., van Merrienboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. Educational Psychology Review, 10, 251-296.
4. Koedinger, K., & Aleven, V. (2007). Exploring the assistance dilemma in experiments with Cognitive Tutors**.** *Educational Psychology Review*. 19(3) 239-264.
5. Celce-Murcia, M., and Larsen-Freeman, D. (1983). The Grammar Book: An ESL/EFL teacher's course. Rowley, Massachusetts: Newbury House Publishers.
6. Master, P. (1994). The effect of systematic instruction on learning the English article system. In T. Odlin (Ed.), *Perspectives on pedagogical grammar,* pp. 229-252. Cambridge: Cambridge University Press.
7. Kurup, M., Greer, J., & McCalla, G. (1992). The Fawlty Article Tutor. Intelligent Tutoring Systems 1992, pp 84-91.
8. Master, P. (1997). The English Article System: Acquisition, Function, and Pedagogy. System. 25,(2) 215-232.
9. Koedinger, K. R., Aleven, V., Heffernan. T., McLaren, B. & Hockenberry, M. (2004). Opening the Door to Non-Programmers: Authoring Intelligent Tutor Behavior by Demonstration. In the Proceedings of 7th Annual Intelligent Tutoring Systems Conference. Maceio, Brazil.
10. VanLehn, K., Siler, S., Murray, C., Yamauchi, T., & Baggett, W.B. (2003). Why do only some events cause learning during human tutoring? Cognition and Instruction, 21(3), 209-249.
11. Fuchs, M., Bonner, M., & Westheimer M. (2006). Focus on Grammar 3: An integrated skills Approach, Third Edition. White Plains, NY: Pearson Education, Inc.
12. Huckin, T. & Olsen L. (1983). English for Science and Technology: A handbook for nonnative speakers. McGraw-Hill, Inc.
13. Ericsson, K. A, & Simon, H. (1984). Protocol analysis: Verbal reports as data. Cambridge, MA: MIT Press
14. Wylie, R. (2007) Are we asking the right questions? Understanding which tasks lead to the robust learning of English articles. Proceedings of the 13th International Conference on Artificial Intelligence in Education (Young Researchers Track).