Talk Like an Electrician: Student Dialogue Mimicking Behavior in an Intelligent Tutoring System

Talk like an Electrician: Student dialogue mimicking behavior in an Intelligent Tutoring System

Natalie B. Steinhauser[1], Gwendolyn E. Campbell[1], Leanne S. Taylor[2], Simon Caine[2], Charlie Scott[2], Myroslava O. Dzikovska[3], and Johanna D. Moore[3⋆]

[1] Naval Air Warfare Center Training Systems Division, Orlando, FL, USA
{natalie.steinhauser,gwendolyn.campbell}@navy.mil
[2] Kaegan Corporation, 12000 Research Parkway, Orlando, FL 32826-2944
Leanne.Taylor.ctr@navy.mil,Simon.Caine.ctr@navy.mil,CScott@kaegan.com
[3] School of Informatics, University of Edinburgh, Edinburgh, United Kingdom
{m.dzikovska,j.moore}@ed.ac.uk

Abstract. Students entering a new field must learn to speak the specialized language of that field. Previous research using automated measures of word overlap has found that students who modify their language to align more closely to a tutor's language show larger overall learning gains. We present an alternative approach that assesses syntactic as well as lexical alignment in a corpus of human-computer tutorial dialogue. We found distinctive patterns differentiating high and low achieving students. Our high achievers were most likely to mimic their own earlier statements and rarely made mistakes when mimicking the tutor. Low achievers were less likely to reuse their own successful sentence structures, and were more likely to make mistakes when trying to mimic the tutor. We argue that certain types of mimicking should be encouraged in tutorial dialogue systems, an important future research direction.

Keywords: Mimicking, Alignment, Intelligent Tutoring System (ITS), Human Computer Interaction (HCI)

1 Introduction

One component of learning a new domain is to learn the "language" of that domain. This includes not only the domain-specific vocabulary, but also the appropriate phraseology and knowledge of how to construct an argument or explanation in that domain. Being able to speak the language of a domain is necessary for effective communication with members of the relevant professional

⋆ This work has been supported in part by US Office of Naval Research grants N000141010085 and N0001410WX20278. We would like to thank our sponsors from the Office of Naval Research, Dr. Susan Chipman and Dr. Ray Perez, and former Research Associates who worked on this project, Katherine Harrison, Elaine Farrow, and Charles Callaway for their contributions to this effort.

community. Students should begin to learn how to "talk like an electrician" (or doctor, or lawyer,etc.) in the classroom, by mimicking the teacher's or tutor's use of domain-specific language. In a computer tutoring context, it is even more important that the student copy the system's use of language, as none of the existing systems are able to understand a full range of natural language input.

It has long been observed that people modify their use of language to correspond more closely with the language used by the person or system that they are communicating with. This basic phenomenon has been studied in many contexts using a variety of different labels and definitions, in particular "alignment" [7], "convergence" [9], "lexical entrainment" [1], and "cohesion" [8].

There is also evidence that the presence of this behavior in student dialogues is positively predictive of measures of student learning. Ward & Litman [8, 9] defined lexical cohesion as the percentage of co-occurrence of individual words within consecutive pairs of dialogue turns, and lexical convergence as the rate of lexical change over a window of 5 to 50 turns. In their curriculum, for students with below average pre-test scores, higher lexical cohesion and convergence scores between the student and the tutor during the tutorial dialogue was predictive of a higher learning gain score. On the other hand, cohesion assessed on pairs of utterances made by the same speaker, whether it was the tutor or the student, was not correlated with learning gain. Ward & Litman concluded that a low level of convergence may indicate that the student is not aligning semantically with the tutor and therefore not learning.

To date, the majority of the research investigating the relationship between alignment and learning gain in computer-based tutoring environments has focused primarily on lexical alignment. Measures of lexical alignment are easy to compute automatically, and it has been theorized that alignment at one level leads to alignment at other levels [7]. In the current study, however, we attempt to extend the previous research by explicitly broadening our definition of linguistic overlap to incorporate features of both lexical and syntactical alignment. We hypothesize that this broader measure should be important in a training context because it reflects the extent to which students use the "language" of a new domain - i.e., not only repeating domain content words, but also organizing those words in meaningful and approriate sentences. In addition, whenever students align at both levels this is more likely to result in utterances that are easy to understand for current computer systems, give state of the art in Natural Language Processing (NLP). In the remainder of the paper we present our measure, which we call "mimicking," and describe our research testing our hypothesis that the amount of mimicking a student produces during a tutoring session will be positively correlated with their learning gain.

2 Method

2.1 Data Collection Environment

The Basic Electronics and Electricity Tutorial Learning Environment (Beetle II)[6] was used for data collection. The Beetle II curriculum of interest in

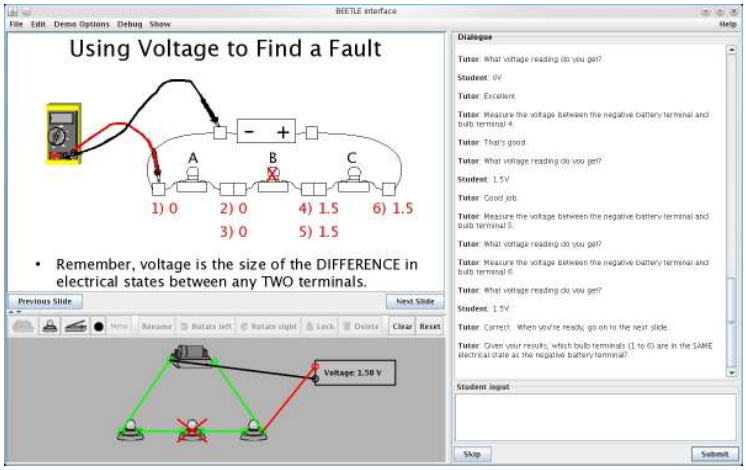

Fig. 1. Screenshot of the BEETLE II system

this study is a lesson on basic electricity and electronics that covers topics such as open and closed paths, voltage reading between components, and finding faults in a circuit with a multimeter. Students took approximately three hours to complete this lesson.

The screen of the BEETLE II system can be seen in Figure 1. It contains lesson material in the form of a self-paced page-turning slide show, a circuit simulator which allowed the students to build and manipulate circuits as a complement to the lesson material, and a chat window where the participants and computer tutor interacted. All interactions with the tutor were typed.

2.2 Procedure

After reviewing the informed consent, participants filled out a demographic questionnaire and took a 22 question pre-test. The participants were then introduced to BEETLE II and given a brief demonstration on the functionality of the learning environment. The students spent the majority of the experimental session working through the lesson materials. During the lesson, the computer tutor instructed the student to read slides, build circuits, and asked the student questions about the material. Every time the student responded to a question, the tutor would provide appropriate feedback.

When the student's answer was correct the tutor would reinforce the answer by either acknowledging that it was correct (e.g., "that's great") or by providing a better way to phrase the answer if the student was right, but not stating the answer in the ideal way (e.g., "Very good. Terminal 1 is connected to terminal 2."). We called the latter a "model better answer" strategy [4]. When the student answered incorrectly, the tutor responded with a a hint to help the student come up with the correct answer on their own. If the student could not get the answer

after three increasingly detailed and specific remediations, the tutor would give the student the answer (e.g., "Almost. Here's the answer. The positive battery terminal is separated by a gap from terminal 3."). We called this a "bottom out".

In about 13% of cases, the system was unable to interpret the student's utterance. In those instances, the system produced an error message indicating that the student was not understood and the reason for misunderstanding, e.g., "I am sorry, I'm having trouble understanding. I didn't understand the word 'power' " [3]. It then asked the student to rephrase their answer, providing a hint depending on the tutoring policy. We will refer to these errors as "uninterpretable utterances." Similar to the case of multiple errors, the system used the "bottom out" strategy if the student made too many uninterpretable utterances.

After the students had completed the lesson, they took a 21 item post-test and filled out a satisfaction questionnaire.

3 Corpus Annotation

The original corpus was comprised of dialogues from forty-one participants. Previous research has shown that the relationship between lexical alignment and learning was strongest for the weakest students. Thus, for this preliminary investigation, we focused on the subset of the corpus that we believed to be most likely to demonstrate an effect of mimicking, the students at the extremes of our distribution. More specifically, we calculated a (normalized) gain score for each student as $\frac{(post-pre)}{1-pre}$. Next, we rank ordered our participants based on this gain score and selected the dialogues from the top ten and bottom nine students. A quick double-check confirmed that, as expected, the high gainers ($M = .75$, $SD = .04$) had a significantly higher learning gain score than the low gainers ($M = .40$, $SD = .17$), $t(17) = 8.67$, $p < .001$.

Of these 19 participants, nine were male and ten were female. Participants' ages ranged from 18 to 25 years with an average age of 20. The final corpus included 770 student turns ($M = 40.5$ per student, $SD = 9.88$).

Our next step was to come up with an operational definition of mimicking that captured the majority of cases of both lexical and syntactical alignment within our corpus, and could be reliably coded by human raters. This is where we were able to rely upon a special feature of our curriculum, which is that many topics are addressed through a series of semi-repetitive questions. For example, in the exercise shown in Figure 1, the students are asked to measure voltage at 4 different points in a circuit. They are then asked a series of questions about the measurements they obtained: "Why did you get the voltage of 1.5 between terminal 1 and the positive battery terminal?", "Why did you get the voltage of 0 between terminal 2 and the positive battery terminal?", and so on. Once an acceptable answer to the first question has been established (either by the student or by the tutor), the student has an opportunity to "mimic" it, i.e., re-use that answer with minor changes for the following questions. For example, in the top left column in Table 1, the student gives a correct answer to the

question. After the tutor acknowledges it as correct, the student uses exactly the same sentence to answer the next question, only modifying it to refer to terminal 6 instead of terminal 5.

Within this framework, we defined mimicking as re-using a complete previous answer, with two minor variations allowed: substituting the component being referenced (e.g., using "bulb A" instead of "bulb B"), and adding or removing negation (e.g., saying "not connected to" instead of "connected to"). Before beginning to code for mimicking, we identified 25 questions where the student has an opportunity to mimic the answer to a previous question. Three independent raters then coded the transcripts, coding each student answer to those questions as either (a) new, (b) a mimic of a previous statement made by the tutor, or (c) a mimic of a previous statement made by the student.[4] Two transcripts were coded by multiple coders to assess inter-rater reliability, which proved to be high (kappa = 0.88).

This way of defining mimicking as a re-use of a statement with only minor changes may seem stringent, but it works well within the context of our curriculum: it reflects strong lexical and syntactic alignment and can be unambiguously recognized by human raters. We return to this in Section 5.

As alluded to above, there are two potential sources of mimicking behavior. First, the students could mimic themselves, by repeating their own previous answers with minor modifications. We refer to this as "self-mimicking". Second, the students can mimic the answers the tutor gives when either the "bottom-out" or "model better answer" strategy is used. We refer to this as "tutor-mimicking".

4 Results

First, we tested our hypothesis that mimicking behavior is positively correlated with learning gains. The overall number of mimicked turns was not significantly correlated with learning gains, $r(18) = -0.17$. The number of self-mimics was also not significantly correlated with learning gains, $r(18) = 0.27$. The number of tutor-mimics was significantly negatively correlated with learning gains, $r(18) = -0.49$, $p < 0.05$.

Given that our results were in direct contradiction to findings from previous research, we extended our coding system to incorporate a feature that the raters had noted during the original coding process – specifically, the fact that sometimes the students' mimicking behavior produced correct answers to new questions, and sometimes it did not. The success or lack of success of each student turn coded as a mimic was determined by the computer tutor's response to that student turn. If the computer tutor accepted the student answer as correct, the mimic was categorized as successful. If the computer tutor was not able to interpret the student answer, or if the computer tutor determined that the student answer was missing essential information or contained one or more

[4] The BEETLE II system was unable to change its phrasing, so only student turns were coded.

	Successful	Unsuccessful
Self-Mimic	**Tutor:** Explain why you got a voltage reading of 0 for terminal 5 and the positive terminal. **Student:** *terminal 5 is connected to the positive terminal of the battery.* **Tutor:** Very good. **Tutor:** Explain why you got a voltage reading of 0 for terminal 6 and the positive terminal. **Student:** *terminal 6 is connected to the positive terminal of the battery.* **Tutor:** Perfect.	**Student:** *There was a gap between terminal 2 and the positive terminal.* **Tutor:** Very good. **Tutor:** Explain why you got a voltage reading of 1.5 for terminal 3 and the positive terminal. **Student:** *There was no gap between terminal 3 and the positive terminal.* **Tutor:** Some of your answer is right, but some is wrong. Please try again.
Tutor-Mimic	**Tutor:** Here's the answer. *The positive battery terminal is separated by a gap from terminal 1.* **Tutor:** Explain why you got a voltage reading of 1.5 for terminal 2 and the positive terminal. **Student:** *The positive battery terminal is separated by a gap from terminal 2.* **Tutor:** That's great.	**Tutor:** Here's the answer. *The positive battery terminal is separated by a gap from terminal 1.* **Tutor:** Explain why you got a voltage reading of 0 for terminal 4 and the positive terminal. **Student:** *the positive battery terminal is separated by a gap from terminal 4.* **Tutor:** Some of your answer is right, but some is wrong. Please try again.

Table 1. Examples of successful and unsuccessful mimics

errors, then the mimic was categorized as unsuccessful. Examples of successful and unsuccessful, self-mimics and tutor-mimics can be found in Table 1.

Once the transcripts were coded for mimicking success, the data were tabulated and summarized (see Figure 2). There was a significant difference between our high and low gainers in the percentage of self-mimicking they produced $t(17) = 2.17$, $p = .05$ and in the percentage of unsuccessful tutor-mimics $t(17) = -3.17$, $p = .01$.

Next, we investigated the relationship between mimicking and uninterpretable utterances. None of the existing dialogue systems are able to interpret the full range of human speech and we have previously shown that high frequency of uninterpretable utterances is negatively correlated with learning gain [5]. We found that overall percentage of mimicking was significantly negatively correlated with percentage of uninterpretables in dialogue ($r = -0.52$, $p = 0.02$), and this correlation was primarily explained by self-mimicking ($r = -0.46$, $p = 0.04$), while

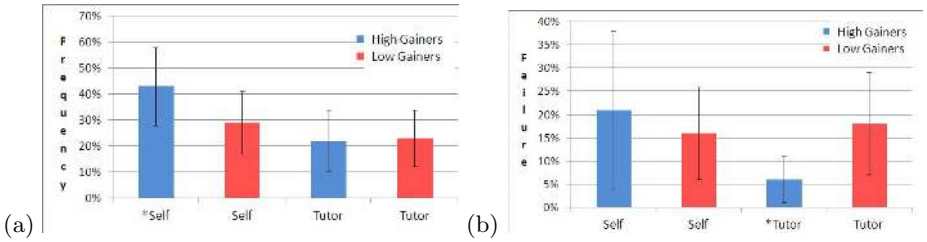

Fig. 2. Assessing Self and Tutor-mimicking between high and low gainers on (a) Frequency and (b) Failures.

tutor-mimics were not significantly correlated with uninterpretables ($r = 0.04$, $p = 0.88$).

Finally, we looked at correlations between the reported overall satisfaction with the system and the amounts of various types of mimics. We found that self-mimicking was correlated with overall satisfaction with the system, $r(18) = 0.52$, $p < 0.05$. However, tutor-mimicking was not significantly correlated with overall system satisfaction, $r(18) = -0.36$, $p > 0.05$.

5 Conclusion

Past research, using a word-by-word method for assessing overlap in two language samples, has shown that the more a student's language converges towards the tutor's language, the higher the student's learning gains, especially among poorer students. In the current study, we moved to a new domain and learning task, and, more importantly, used a different measure of alignment (which we call mimicking), focusing on an amalgamation of lexical and syntactical alignment. We initially hypothesized, like past research, that student mimicking of the tutor would yield higher learning gains and improve communication with the system.

Our results produced a more complex pattern between the variables than was found in previous research. Students with the highest learning gains were more likely to mimic themselves and were more satisfied with the system. It appeared that these students found a strategy for responding to the tutor's questions that was successful and then stuck with it as much as possible. On the other hand, students with the lowest learning gains were less likely to mimic themselves, less able to successfully mimic the tutor and were less satisfied with the system. Moreover, for all students, the more they engaged in self-mimicking, the more successful they were in communicating with the tutor.

These results suggest that it may be advantageous to encourage certain types of mimicking behaviors in a tutorial setting and particularly with an ITS. Mimicking will help students to "talk like an electrician" (i.e., learn the proper way of speaking in the domain) and will help them to be understood by the system, which should make for better dialogue and a more enjoyable experience with the system. The current system incorporated features designed to facilitate tutor-mimicking. For example, the "bottom-out" and "model better answer" strategies

were intended to provide patterns that students could imitate. Our results indicate, however, that self-mimicking is a more important predictor of learning gain. The best strategies to encourage self-mimicking are an open question for future work.

We used the special property of our curriculum, namely, the presence of semi-repetitive questions, to account for syntactic alignment without the need of syntactic parsing based on surface properties of student answers. Our definition was based on phenomena frequently observed in our corpus, and designed to achieve high inter-rater reliability. It would be possible to relax it slightly, in particular, to allow for use of pronouns and discourse connectives, and we are considering extending our analyses to cover those cases. The results also need to be replicated with different domains and curricula. However, in absence of similar questions, automated NLP tools would have to be used. Possibilities include using syntactic parsers or other automatically computable measures of cohesion (e.g. those used in Coh-Metrix [2]).

It is also possible that the importance of mimicking and the ease or difficulty of mimicking successfully may be affected by the quality of NLP and the nature of the domain. If the system error rate decreases, the percentage of successful mimics may increase and the relationship between successful mimicking and learning gain may change. Thus, these results should be re-examined as advances are made in the state of the art in natural language interpretation.

References

1. Brennan, S.E.: Lexical entrainment in spontaneous dialog. In: Proceedings of the 1996 International Symposium on Spoken Dialogue. pp. 41–44 (1996)
2. D'Mello, S.K., Dowell, N., Graesser, A.C.: Cohesion relationships in tutorial dialogue as predictors of affective states. In: Proceedings of AIED-2009. pp. 9–16 (2009)
3. Dzikovska, M.O., Callaway, C.B., Farrow, E., Moore, J.D., Steinhauser, N.B., Campbell, G.C.: Dealing with interpretation errors in tutorial dialogue. In: Proceedings of SIGDIAL-09. London, UK (Sep 2009)
4. Dzikovska, M.O., Campbell, G.E., Callaway, C.B., Steinhauser, N.B., Farrow, E., Moore, J.D., Butler, L.A., Matheson, C.: Diagnosing natural language answers to support adaptive tutoring. In: Proceedings of the 21st International FLAIRS Conference (2008)
5. Dzikovska, M.O., Moore, J.D., Steinhauser, N., Campbell, G.: The impact of interpretation problems on tutorial dialogue. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (2010)
6. Dzikovska, M.O., Moore, J.D., Steinhauser, N., Campbell, G., Farrow, E., Callaway, C.B.: Beetle II: a system for tutoring and computational linguistics experimentation. In: Proceedings of the ACL-2010 demo session (2010)
7. Pickering, M.J., Garrod, S.: Toward a mechanistic psychology of dialogue. Behavior and Brain Sciences 27, 169–226 (2004)
8. Ward, A., Litman, D.: Cohesion and learning in a tutorial spoken dialog system. In: Proceedings of 19th International FLAIRS Conference (2006)
9. Ward, A., Litman, D.J.: Dialog convergence and learning. In: Proceedings of the 13th International Conference on Artificial Intelligence in Education (2007)