# Dialog Convergence and Learning

Arthur WARD and Diane LITMAN
*Learning Research and Development Center, University of Pittsburgh*

**Abstract.** In this paper we examine whether the student-to-tutor convergence of lexical and speech features is a useful predictor of learning in a corpus of spoken tutorial dialogs. This possibility is raised by the Interactive Alignment Theory, which suggests a connection between convergence of speech features and the amount of semantic alignment between partners in a dialog. A number of studies have shown that users converge their speech productions toward dialog systems. If, as we hypothesize, semantic alignment between a student and a tutor (or tutoring system) is associated with learning, then this convergence may be correlated with learning gains. We present evidence that both lexical convergence and convergence of an acoustic/prosodic feature are useful features for predicting learning in our corpora. We also find that our measure of lexical convergence provides a stronger correlation with learning in a human/computer corpus than did a previous measure of lexical cohesion.

**Keywords.** Intelligent Tutoring, Learner Modeling, Discourse Analysis

## 1. Introduction

Human tutors have been shown to produce significantly larger learning gains than classroom instruction [1]. Because these human tutors teach using natural language dialog, many Intelligent Tutoring System (ITS) researchers are investigating the addition of natural language interfaces to their tutoring systems [2, 3]. To help inform tutoring system design, many researchers search for aspects of natural language tutoring dialogs which might be associated with learning. Various types of dialog features have been investigated [4–6], many of which have demonstrated correlations to learning. Often, however, these features are difficult to implement in an ITS because they require hand coding of dialog transcripts. Shallow acoustic/prosodic [7], as well as other dialog features have been investigated which are automatically computable, but which may be less strongly related to learning depending on their linguistic sophistication [8]. As a result, there remains a need in the tutoring community for automatically computable dialog features which are also predictive of learning.

Work outside the tutoring community has suggested links between certain dialog characteristics and the amount of understanding that develops between dialog partners. In particular, Pickering and Garrod's Interactive Alignment Model [9] suggests a link between convergence and semantic alignment. In this paper we use the term "convergence" to mean when two dialog partners change some aspect of

their speech to be more similar to each other. We use the term "alignment" to mean when the internal representations they create become more similar to each other. "Priming," as described below, is one mechanism thought responsible for alignment and convergence.

The Interactive Alignment Model posits that each dialog partner processes speech in several levels. The speech signal is unpacked into low level phonetic and phonological representations, into lexical and syntactic representations, and so upward until high level semantic and situation model representations are determined. In this way, a number of internal representations of the incoming speech become active during processing. If speech is then produced while they are still active, these representations are more likely to be used than others that are less active. This priming process leads to alignment between the internal representations used at each level by the two dialog partners, and also to the convergence of their observable speech features. Priming is also thought to link neighboring levels of representation, so that alignment at one level increases the tendency to align at neighboring levels. We hypothesize that lexical and acoustic/prosodic (a/p) convergence between tutor and student is linked with alignment of their semantic representations, which may in turn be linked to learning.

Researchers have found that users will converge toward a dialog system on several lexical and a/p features. For example Coulston et al. [10], have found evidence for the convergence of spoken amplitude (loudness) toward that of a dialog agent. Bell et al. [11] have found that users will converge their speech rate toward that of a dialog system, and Brennan [12] has found that users will converge toward the lexical choices made by a system.

In section 2, we describe corpus measures of convergence which we developed in previous work [13]. In that work we showed that convergence was present in our corpus of tutoring dialogs with a human tutor. Here we show that our measures are also useful predictors of learning. In particular, we will show that our measure of lexical convergence predicts learning for the below mean pre-testers in both our human/human (hh) and our human/computer (hc) corpora. Lexical convergence is also a significant predictor of learning for all students in the hc corpus. One of our measures of a/p convergence will also be shown to predict learning among the high pre-testers in our hc corpora. Finally, we compare our measure of lexical convergence to our previous measure of lexical cohesion [14], and find convergence to have a stronger and more significant correlation with learning in our hc corpus.

## 2. Measuring Convergence

In [13] we built on previous work by Reitter et al. [15] to develop measures of lexical and a/p convergence which we applied to a corpus of tutorial dialogs. Both measures, lexical and a/p, proceed in the same general way. They both first locate a prime in tutor utterances, then define the next $N$ student turns following this prime as a response window. For each distance $d$ from the prime within the response window, they record the value of a response variable. After these data points have been collected for every prime in the corpus, linear regression is used to model the interaction between distance from the prime and the value

of the response variable. Following Reitter, a negatively sloped regression line is interpreted to mean an increased value of the response variable immediately following the prime, which then declines back toward it's global mean.

For the measure of lexical priming, the prime was defined to be the occurrence of some word $w$ in a tutor utterance. Each word in the tutor utterance is treated as a potential prime, except, as described in [13], words which we classified as having had no alternative synonym. This adjustment was designed to make our measure better reflect priming's effect on lexical choice. In lexical priming, the response variable is the re-occurrence of the prime word in any of the following $N$ student turns. For example, if the prime word was used again in the third student utterance following a prime, the data point [3,1] would be collected. In [13] we collected data sets for window sizes of 5, 10, 15 and 20 student turns, and fit lines to each of them using linear regression. The lexical slopes at the largest three window sizes were negative and significantly different than zero, which we argued was evidence for lexical priming effects.

We also looked for priming effects in six acoustic/prosodic (a/p) features: max, mean and min RMS (loudness) and max, mean and min f0 (pitch). For these a/p measures the prime was located wherever the tutor's value for the a/p feature in question was more than one standard deviation above the tutor's global mean. As with lexical priming, a response window was defined to be the next N student turns following the tutor's prime. The value of the a/p feature was recorded for each utterance in this window. For example, if the a/p feature in question was maxRMS (ie, maximum "loudness" in the turn), and the maxRMS value for the first student turn following the prime happened to be "1280," then the data point [1, 1280] would be collected. Linear regression produced slopes that were negative and significantly different from zero for max RMS at window sizes of 25 and 30, and for mean RMS at a window size of 30 student turns. Slopes were positive and significantly different from zero for min f0 at window sizes of 15, 20, 25 and 30 student turns. Significance thresholds were adjusted for multiple comparisons.

Both the lexical and a/p priming measures were validated by comparing their performance on naturally ordered vs. randomized data. The measures produced significant slopes on the naturally ordered data, but not on the randomized data. We argued that the lack of false positives on randomized data was evidence for the reliability of our measures. In [13], we fit regression lines to an entire corpus to show that priming effects could be detected in tutorial dialog. In the current work, we use the same method, but fit lines individually to each student in the corpus. We then show that the slope of the student's fitted regression line is a useful feature for predicting learning gains.

## 3. The Corpora

Our training set is the same corpus of tutoring sessions used in [13] to develop our measures of priming. In these tutoring sessions, a student first takes a pre-test to gauge knowledge of relevant physics concepts, then reads instructional material about physics. Following this, the human tutor presents a problem which the student answers in essay form. The tutor then examines this essay, identifies flaws

in it, and engages the student in a tutorial dialog to remediate those flaws. The student then reworks the essay, and the cycle repeats until the tutor is satisfied that all important points are covered. Each student finished up to ten problems. After the final problem, each student took a post-test, and learning gains were calculated. The resulting corpus contains sessions with fourteen students, totaling 128 dialogs, 6,721 student turns and 5,710 tutor turns. A mean pre-test split produces eight low and six high pre-testers.

Our testing corpus is a similar collection of tutoring dialogs, but collected using the ITSPOKE [16] spoken dialog tutoring system. ITSPOKE is a speech-enabled version of the WHY2-Atlas intelligent tutoring system [17]. The ITSPOKE system also teaches qualitative physics, and engages the student in the same cycle of dialog and essay revision as did the human tutor. In this corpus twenty students engaged in up to five tutorial dialogs each, resulting in a corpus of 95 dialogs, 2,335 student turns and 2,950 tutor turns. A pre-test split produces thirteen low and seven high pre-testers.

Our two corpora were similar in many respects, such as having similar subject matter and a similarly structured tutoring cycle. However, they were also different in two ways that may be relevant to the current study. First, the average number of student words per turn was higher with the human tutor. Students in the hh corpus averaged 5.21 words per turn, students in the human/computer (hc) corpus averaged 2.42 words per turn. Second, student utterances in the human tutor corpus seem to contain a broader range of words and have more complex syntactic structure. Student utterances in the computer tutor corpus, on the other hand, seem to be much more terse, containing more "keyword only" answers.

## 4. Finding Which Features Predict Learning

We want to find which, if any, of the measures of priming described in section 2 are associated with learning. To do this, we allow an automatic feature selection algorithm to select predictors on the hh corpus, then we test the validity of the selection by fitting a new linear model, which contains the selected features, to the hc corpus. If the features are significant predictors also in the second corpus, we take that to be evidence that the algorithm has selected useful features.

The automatic feature selection algorithm we use is "stepwise regression." It starts with an empty linear model, then adds and removes one variable at a time while attempting to minimize the Akaike Information Criteria (AIC) [18].

We first select features using all students, then separately for students with above-mean and below-mean pre-test scores. We do this because in previous work [14] our high and low pretesters responded differently to lexical features of the dialog. In addition, we will select features starting from several different initial feature sets. First we will allow the feature selection algorithm to choose among all ten features described in section 2. Then, we run feature selection again within each category. That is, we start with all features, then with only lexical features, then a/p features. We do this for two reasons. First, it will help avoid bias in feature selection, given that stepwise regression can produce different sets of predictors given slightly different starting points. Second, we are interested in

how the a/p features perform in isolation, because the lexical features depend on reliable speech transcriptions, which are often not available at runtime.

This approach resulted in nine runs of stepwise regression (3 student groupings x 3 initial feature sets). These nine runs produced three significant linear models. The remaining runs produced either empty models or models containing only "pretest" as a predictor. We test the usefulness of the selected predictors by using them to fit new models on our hc (human-computer) corpus.

| | | | | | | **Pretest Only** | | **Full Model** | |
|---|---|---|---|---|---|---|---|---|---|
| | **N** | **Student Group** | **Selected Features** | | | **Model pVal** | **Adj $R^2$** | **Model pVal** | **Adj $R^2$** |
| | | | **Inter** | **preTest** | **lex.w20** | | | | |
| 1h | 14 | All hh | **0.484*** ** | **0.652*** | 3.422 | **0.012** | 0.368 | **0.017** | 0.434 |
| 2h | 8 | hh low pre | 0.331 | **1.267*** | **7.901*** | 0.29 | 0.047 | **0.044** | 0.597 |
| 1c | 20 | All hc | **0.315**** | **0.583**** | **-11.099**** | **0.044** | 0.162 | **0.003** | 0.431 |
| 2c | 13 | hc low pre | **0.866**** | -1.075 | **-17.361**** | 0.796 | -0.08 | **0.004** | 0.589 |

**Table 1.** Lexical features selected on hh data (models 1h & 2h), tested on hc data (1c & 2c) Significance codes: p < .05: *; p < .01: **; p < .001: ***

| | | | | | | **Pretest Only** | | **Full Model** | |
|---|---|---|---|---|---|---|---|---|---|
| | **N** | **Student Group** | **Selected Features** | | | **Model pVal** | **Adj $R^2$** | **Model pVal** | **Adj $R^2$** |
| | | | **Inter** | **preTest** | **meanRMS.w20** | | | | |
| 3h | 6 | hh hi pre | 0.255 | **0.998*** | -0.015 | **0.024** | 0.694 | **0.024** | 0.860 |
| 3c | 7 | hc hi pre | **0.315*** | 0.816 | **0.026**** | 0.089 | 0.363 | **0.034** | 0.725 |

**Table 2.** a/p features selected on hh data (model 3h),tested on hc data (3c)

The first of these three significant models was selected when starting with all features and all students. This is shown as model 1h in Table 1 (model numbers are shown in column one). In table 1 individual coefficients with p-values below .05 are shown in bold, with the level of significance indicated by asterisks (see caption). The model's selected features and their values can be read under the "Selected Features" columns. Model 1h can be read as: $posttestscore = .48 + .65 * pretest + 3.42 * lex.w20$. Lex.w20 is the slope of the lexical response line, fitted to each student using a window size of twenty. This model predicts that post-test score will increase .65 points for each point increase in pre-test score. Also, post-test score is predicted to increase by 3.42 points for each point increase in the lexical slope. Pre-test score is a reasonable selection, because it is correlated with post-test score in our human-human corpus (R = .64, pVal = .012). Lex.w20 is not itself significant in this model, but was selected because it improved the fit of the model by more than the AIC penalty for additional factors. The model p-value and adjusted $R^2$ given in the "Pretest Only" columns are for a model containing only an intercept and pre-test. The same numbers given in the "Full Model" columns are for a model which also includes the third predictor, in this case "lex.w20." The additional value of the lex.w20 predictor can be seen by comparing these two sets of numbers. In every case, the model's adjusted $R^2$ is larger for the full model. In almost every case, the model's p-value is better, as well.

The second significant model was selected when starting with all features and the low pre-test students, and is shown as model 2h of Table 1. This model contains the same features as were selected for model 1h, however in this model lex.w20 becomes individually significant.

Next, we test the features selected on the hh corpus by fitting them to our hc corpus. Model 1c of Table 1 shows the result of fitting the features of model 1h to the set of all hc students. The lex.w20 feature becomes highly significant here. Model 2c of Table 1 shows how the features of model 2h fare when fitted to the low pre-test students in the hc corpus. Lex.w20 is individually significant in this model, as it was in the hh corpus.

As described above, we also started feature selection from each category of features separately. Only one significant model was found in these runs, which is shown as Model 3h in Table 2. When starting with a/p (acoustic/prosodic) features and the hh high pre-testers, stepwise regression selects pre-test score and meanRMS.w20. MeanRMS.w20 is not individually significant when initially selected on the hh data. Model 3c in Table 2 shows the results of fitting the features from Model 3h to the hc data. The model as a whole remains significant on the hc data, while meanRMS.w20 becomes individually significant.

We have seen that features selected using only the human-human corpus perform well when fitted to a very different corpus of human-computer dialogs. This suggests that they may be genuinely useful indicators of learning in tutoring dialog. It is interesting to note, however, that the values fitted to these features differ widely between the corpora. When it is significant, the lexical priming feature "lex.w20," is fitted with positive coefficients in the human-human data, but with negative coefficients in the human-computer data. The meanRMS.w20 feature also is fitted with different signs in the two corpora, although in the hh corpus it is not individually significant[1].

Note that these features represent the slope of a line fitted to the student's response after a prime. A negative coefficient fitted to these features means that the student learns more as convergence increases and the slope becomes more negative. The negative lexical coefficients fitted to the hc corpus thus fit the predictions of the Interactive Alignment Model.

### *4.1. Lexical Cohesion vs. Convergence*

As mentioned above, the usefulness of splitting our data by pre-test scores was suggested by previous work with a measure of lexical cohesion [14]. We now use that measure of lexical cohesion as a benchmark for comparing our current results. In [14] we measured cohesion by counting the number of cohesive ties between student and tutor, where a cohesive tie was defined as lexical overlap between consecutive speakers. Table 3 reproduces some of the results from that study. Rows two and three in Table 3 divide the students into above-mean and below-mean pre-test categories, exactly as in the present study. The center column shows the partial

[1]To test whether multicollinearity between window sizes might be preventing the selection of predictors which didn't switch sign between corpora, we also ran individual models for each predictor at each window size. All predictors kept their sign across different window sizes. We thank an anonymous reviewer for this suggestion.

correlation of cohesion and post-test, controlling for pre-test, for each group of students. The right column shows the significance of the correlation. Table 4 has the same layout, showing results for the partial correlation of the lex.w20 measure and post-test score, controlling for pre-test score. Both metrics are run on the same data set. Note that the significant lexical convergence correlations in Table 4 have consistently larger absolute magnitudes than the cohesion correlations in Table 3. Also, the convergence measure produces a significant correlation for the group of all students. This suggests that lexical convergence is a better predictor of learning than our previous measure of lexical cohesion.

| **HC Data** | R | P-Value |
|---|---|---|
| All Students | 0.431 | 0.058 |
| Low Pretest | 0.676 | **0.011** |
| High Pretest | 0.606 | 0.149 |

**Table 3.** Lexical Cohesion

| **HC Data** | R | P-Value |
|---|---|---|
| All Students | -0.599 | **0.005** |
| Low Pretest | -0.809 | **0.001** |
| High Pretest | -0.437 | 0.327 |

**Table 4.** Lexical Convergence

## 5. Discussion and Future Work

The differences in coefficient polarity shown in Tables 1 and 2 are intriguing, and not yet explained. In section 3 we described several differences between the corpora which might be relevant. First the difference in student turn length between the hh and hc corpora may mean that our windows, while having the same size in turns, may be different in the number of words or amount of time they contain. Reitter et al. [15], on whose work we modeled our measures of convergence, experimented with defining response windows as spans of time. Possibly defining our windows similarly would remove the polarity differences between the corpora.

The other difference mentioned in section 3 was that there seems to be a difference in content, with the hh corpus containing more non-domain specific words. The difference in lexical polarity may then be partly explainable by differences in the words being used. Perhaps convergence is only positively correlated with learning when measured with domain-relevant tokens.

We have shown that measures of lexical and a/p convergence are useful features for predicting learning in two corpora of tutoring dialogs. When both are applied to the same corpus, our measure of lexical convergence also provides better p-values and larger correlations than our previous measure of lexical cohesion. Our other measure of convergence, which uses the "mean RMS" feature, was useful in predicting learning among our high pre-testers, and possesses the additional benefit of not requiring (machine or manual) transcription.

The corpora used in this study are our only two for which a/p features are currently available. Work is underway to generate a/p features for several additional corpora. Further experiments may help resolve the discrepancy in polarity of results reported here, by determining if fitted coefficients remain negative on future hc corpora. If experiments on these additional corpora are successful, we hope to include real-time measures of convergence in the ITSPOKE student model. A low or declining level of convergence may help indicate when the student is not aligning semantically with the tutor, and perhaps not learning.

## 6. Acknowledgments

This research is supported by the NSF (0325054), and by an Andrew Mellon Predoctoral Fellowship. We gratefully thank Tessa Warren for advice on this project, and also Joel Tetreault and the ITSPOKE group for many helpful comments.

## References

[1] B. Bloom. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13:4–16, 1984.

[2] M. Evens and J. Michael. *One-on-one Tutoring By Humans and Machines*. Erlbaum, 2006.

[3] Heather Pon-Barry, Karl Schultz, Elizabeth Owen Bratt, Brady Clark, and Stanley Peters. Responding to student uncertainty in spoken tutorial dialogue systems. In *International Journal of Artificial Intelligence in Education (IJAIED) Special Issue "Best of ITS 2004"*, volume 16, pages 171–194, 2006.

[4] A. Graesser and N. Person. Question asking during tutoring. *American Educational Research Journal*, 31:104–137, 1994.

[5] C. Rosé, D. Bhembe, S. Siler, R. Srivastava, and K. VanLehn. The role of why questions in effective human tutoring. In *Proceedings of AI in Education*, 2003.

[6] Mark Core, Johanna Moore, and Claus Zinn. The role of initiative in tutorial dialogue. In *10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, April 2003.

[7] K. Forbes-Riley and D. Litman. Correlating student acoustic-prosodic profiles with student learning in spoken tutoring dialogues. In *Proceedings Interspeech-2005/Eurospeech*, 2005.

[8] Kate Forbes-Riley, Diane Litman, Mihai Rotaru Amruta Purandare, and Joel Tetreault. Comparing linguistic features for modeling learning in computer tutoring. In *Proceedings 13th International Conference on Artificial Intelligence Education (AIED)*, Los Angeles, Ca., 2007.

[9] Martin J Pickering and Simon Garrod. Toward a mechanistic psychology of dialogue. In *Behavioral and Brain Sciences*, volume 27, 2004.

[10] R. Coulston, S. Oviatt, and C. Darves. Amplitude convergence in children's conversational speech with animated personas. In *Proceedings of the 7th International Conference on Spoken Language Processing*, 2002.

[11] Linda Bell, Joakim Gustafson, and Mattias Heldner. Prosodic adaptation in human-computer interaction. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 03)*, Barcelona, Spain, 2003.

[12] Susan E. Brennan. Lexical entrainment in spontaneous dialog. In *International Symposium on Spoken Dialog*, pages 41–44, 1996.

[13] Arthur Ward and Diane Litman. Measuring convergence and priming in tutorial dialog. In *Technical report TR-07-148*, University of Pittsburgh, 2007.

[14] A. Ward and D. Litman. Cohesion and learning in a tutorial spoken dialog system. In *Proceedings of the 19th International FLAIRS Conference*, pages 533–538, 2006.

[15] David Reitter, Johanna Moore, and Frank Keller. Priming of syntactic rules in task-oriented dialogue and spontaneous conversation. In *Proceedings of the 28th annual Conference of the Cognitive Science Society*, Vancouver, 2006.

[16] D. Litman and S. Silliman. ITSPOKE: An intelligent tutoring spoken dialogue system. In *Companion Proc. of the Human Language Technology Conf: 4th Meeting of the North American Chap. of the Assoc. for Computational Linguistics*, 2004.

[17] K. VanLehn, P. Jordan, C. Rose, D. Bhembe, M. Boettner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, and Srivastava R. The architecture of why2-atlas: A coach for qualitative physics essay writing. In *Proc. 6th Int. Conf. on Intelligent Tutoring Systems*, pages 158–167, 2002.

[18] Hirotugu Akaike. A new look at the statistical model selection. *IEEE Transactions on Automatic Control*, AC-19(6):716–723, 1974.