Using Natural Language Processing to Analyze Tutorial Dialogue Corpora Across Domains and Modalities

Diane LITMAN [a], Johanna MOORE [b], Myroslava O. DZIKOVSKA [b] and Elaine FARROW [b]

[a] *University of Pittsburgh, Pittsburgh, PA*
[b] *University of Edinburgh, Edinburgh, Scotland*

Abstract. Our research goal is to investigate whether previous findings and methods in the area of tutorial dialogue can be generalized across dialogue corpora that differ in domain (mechanics versus electricity in physics), modality (spoken versus typed), and tutor type (computer versus human). We first present methods for unifying our prior coding and analysis methods. We then show that many of our prior findings regarding student dialogue behaviors and learning not only generalize across corpora, but that our methodology yields additional new findings. Finally, we show that natural language processing can be used to automate some of these analyses.

Keywords. tutorial dialogue, intelligent tutoring, natural language, discourse

1. Introduction

One major difference between human tutors and current computer tutors is that only human tutors participate in unrestricted natural language dialogue with students. The development of automated tutorial dialogue systems has thus emerged as an important research topic in the field of intelligent tutoring systems. Researchers have hypothesized that giving computer tutors the ability to engage in natural language dialogue is one approach for potentially closing the current performance gap between human and computer tutors, with respect to increasing student learning.

Unfortunately, building a tutorial dialogue system is not at all straightforward. Recent work has begun to try to empirically determine how to make such tutorial dialogue systems more effective, by examining how specific student and tutor dialogue behaviors correlate with learning (e.g., [1,2,3]). In our own prior work, for example, we analyzed dialogue behaviors and learning using corpora of computer and human tutoring spoken dialogues in the conceptual mechanics domain [4,5], as well as human tutoring typed dialogues in the electricity domain [6]. However, because researchers – including ourselves in these prior projects - differ with respect to both how they code their dialogue data, and how they statistically analyze it, it is often difficult to evaluate the generality of specific findings. The goal of our current work is to determine whether our prior findings and methods can be generalized across our two sets of previously collected corpora, given

their differing dialogue modalities (spoken versus typed), tutoring domains (mechanics versus electricity in physics), and tutor types (computer versus human).

Section 2 describes our corpora and prior results. Section 3 maps our existing student dialogue annotations to a common representation, and examines the use of natural language processing to automatically create the annotations. Section 4 presents our method for statistically analyzing the uniformly coded data, then uses this method to develop predictive models of learning. Section 5 discusses our results, which suggest that findings and methods can be generalized across tutorial dialogue systems, and that automatically created annotations can yield similar results as those created by humans.

2. Dialogue Corpora and Prior Results

Our spoken tutoring data is in the domain of conceptual mechanics and consists of a corpus of 100 dialogues from 20 students interacting with the ITSPOKE computer tutor. ITSPOKE (Intelligent Tutoring SPOKEn dialogue system) is a speech-enabled version of the text-based Why2-Atlas [7] system. The data was collected during a study comparing spoken versus typed human and computer tutoring [4].

Our corpus of typed tutoring data is in the domain of basic electricity and electronics, and consists of 60 dialogues from 30 students interacting with one of three human tutors. This corpus is being used to inform the design of the BEETLE (Basic Electricity and Electronics Tutorial Learning Environment) tutorial dialogue system. The corpus was collected during a study comparing the use of different forms of tutor questions (open-ended versus short answer versus multiple choice) [6].

When collecting both corpora, students were given a multiple-choice pretest before tutoring, followed by a (non-identical) multiple-choice posttest after tutoring. These tests were used to compute measures of student learning such as learning gain. Also, both corpora were automatically or manually annotated with respect to a variety of tagsets. These tagsets were used to construct quantitative measures of tutor and student dialogue behaviors, which were then examined for correlations with learning. For our current generalization study, we focus only on annotations of student dialogue behaviors that yielded significant correlations with learning in prior ITSPOKE and/or BEETLE studies.

Unit tags were used to automatically compute simple counts of linguistically meaningful units such as number of *Words* or *Turns*. While various measures of dialogue interactivity based on such tags (e.g., the percentage of dialogue produced by the student) positively correlated with learning in BEETLE [8], similar measures of student language production did not correlate with learning in ITSPOKE [4].

Content tags were used to identify dialogue containing information relevant to the domain topic. In ITSPOKE, a lexical item was automatically tagged as a *DomainConcept* whenever it was in an online physics glossary; measures based on these tags (e.g., number of content words, number of content-rich turns) positively correlated with learning [9]. In BEETLE, a human annotator partitioned all dialogue into a set of non-overlapping segments tagged as either *Content, Management, Metacognition, or Social*; the amount of dialogue in the content category positively correlated with learning [8].

Several dimensions underlying a number of theories regarding how student events lead to learning have been manually annotated in both corpora. In BEETLE, student contributions were tagged along a set of five dimensions identified in the DeMAND Cod-

ing System [10].[1] For example, student statements were coded for **Depth** (*Yes, No*) and **Novelty** (*New, Old, Other*), with novelty shown to be positively correlated with learning [8]. Two other DeMAND dimensions (Accuracy and Doubt) will be described below. In ITSPOKE, student "Answers" were coded as *Shallow, Deep, Novel*, and a number of correlations identified. In computer tutoring the presence of student turns that displayed reasoning positively correlated with learning, while in human tutoring the introduction of a new concept positively correlated with learning [5].

Accuracy with respect to the tutoring topic has also been annotated. ITSPOKE data has been tagged as either *Correct*, *Incorrect*, *PartiallyCorrect*, or *Can'tAnswer*. Although this tagging was done automatically by ITSPOKE, accuracy was later manually tagged by a paid annotator to remove noise. BEETLE dialogue was tagged as *Correct*, *Incorrect*, *SomeErrors* or *SomeMissing*; this tagging was done by the human tutor in real time. Accuracy was shown to be positively predictive of learning in BEETLE [8], but did not correlate with learning in ITSPOKE [12].

Finally, tags relating to student affects/attitudes have been manually annotated. BEETLE has been annotated with respect to signs of **Doubt** (*Yes, No*), and doubt was used in conjunction with accuracy to show that "impasses" (wrong, or right but uncertain) negatively correlated with learning [10]. ITSPOKE has been annotated with respect to "Certainness" (*Certain, Uncertain, Neutral, Mixed*) and "Frustration" (*Frustrated, Non-Frustrated*), with neutral student dialogue negatively correlated with learning [12].

3. Creating a Common Research Framework

The results summarized in the prior section suggest that in tutorial dialogue the quality of student responses is particularly predictive of learning. Of particular note are our results regarding domain content and novelty, as positive correlations with learning were found for both ITSPOKE and BEETLE corpora. However, because the specific annotations, measures of learning, and statistical approaches used to obtain our prior results sometimes differed, it is difficult to know whether the findings that seemingly generalized are actually the same. Furthermore, additional findings might also have generalized, if the corpora had been analyzed more uniformly.

To create such uniformity, the first stage of our research involves mapping related but non-identical aspects of our annotation schemes to identical tagsets, or to common higher level theoretical constructs. We take the first approach to automatically and uniformly retag our corpora with respect to units and domain content. We take the second approach for the DeMAND tag dimensions described above, using the theoretical constructs and the dimensions-to-construct mappings proposed by [10] as our starting point. A similar approach has been used by the dialogue community, e.g., the DAMSL tagset of dialogue acts (www.cs.rochester.edu/research/speech/damsl/) was developed to provide a common abstraction that enabled sharing of task-oriented dialogue corpora, but that remained compatible with the original tagsets used to annotate such corpora.

With respect to unit tags, our current study focuses only on annotating *words*, which are automatically identified using a *tokenization* program. While word tokenization might seem straightforward (e.g., determine word boundaries using whitespace), it

[1] These dimensions were hypothesized to be sufficient for describing and differentiating five student-oriented theoretical constructs which could influence tutoring system design [10], e.g. impasses [11].

is actually more complex (e.g., contractions can be counted as one word or two). The original BEETLE and ITSPOKE tokenizers in fact made different decisions regarding contractions and several other issues. For our current work, we agreed on a set of uniform conventions, implemented a new word tokenizer, then retokenized the corpora.

With respect to content tags, as in our prior ITSPOKE work [9], we automatically tag words corresponding to domain-specific concepts using an online list of dictionary entries from *Eric Weisstein's World of Physics* (http://scienceworld.wolfram.com/physics). While content tagging in BEETLE had previously been done manually, in our current study we instead use this dictionary to automatically tag both the ITSPOKE and BEETLE corpora. This allows us to better determine whether content results do indeed generalize across ITSPOKE and BEETLE, and to examine whether our prior BEETLE results change after moving from manual to automatic tags. In addition, since a glossary tailored to the electricity domain was available for BEETLE, we use this glossary to create a second content tagging of the BEETLE data. This allows us to examine whether results change when using a generic versus a specialized domain dictionary.

In our prior work, a word was tagged as a *domain concept* if there was a corresponding single word entry in the online dictionary. Because of this simple approach, some domain words in our corpus (e.g., "vectors", "frictional") were not tagged as content, because they did not exactly match the relevant dictionary entry (e.g., "vector", "frictional force"). To allow matches with both the singular-noun word forms as well as the multi-word entries in our dictionary, we now preprocess both our corpora and our dictionary using natural language processing. We first perform *word tokenization*, as described above. Next, we use a *stop word list* to eliminate high-frequency words, because tokenizing a dictionary entry like "Newton's Law of Cooling" adds the stop word "of" to the domain dictionary. Finally, we use *stemming* to map various morphological variants of a word to a common stem. However, while stemming enables words from a corpus like "forces" to correctly match a dictionary entry like "force", stemming often introduces other types of incorrect matches. To examine whether stemming adds value in our experiments, we create content tags both with and without our stemming preprocessor. We use the Natural Language Toolkit (NLTK, http://www.nltk.org), a suite of software modules for analyzing text with Python, to implement our stemmer and other preprocessors.

For our remaining tags (representing the DeMAND dimensions depth, novelty, accuracy, and doubt), we map the existing BEETLE and ITSPOKE tags and tag values to learning theory constructs, using the dimensions-to-construct mappings in Table 1. The first and second columns list the subset of constructs and BEETLE mappings from [10] that are relevant to the current study; the third column lists a new set of ITSPOKE mappings needed for the current study, given the related but non-identical ITSPOKE tagsets for the DeMAND dimensions. The first two rows of the table show that two common theoretical learning constructs for analyzing student contributions, namely the application of cognitive effort (e.g., [13]) and the construction of new knowledge (e.g., [14]), map directly onto the depth and novelty tagging dimensions, respectively. In contrast, the last two rows show that two other constructs, namely student impasses (contributions that are either expressed with uncertainty but are correct, or contributions that are wrong) (e.g., [11]) and accountable talk (contributions that both involve effort and are correct) (e.g, [15]) instead involve mapping particular tag values from multiple dimensions.[2]

[2] A more elaborated ITSPOKE mapping, where certainness also plays a role for incorrect answers, and different types of impasses are ranked by severity, is presented in [16].

Table 1. Mapping of Original Tags and Values to Learning Theory Constructs

Construct	BEETLE	ITSPOKE
Effort	Depth=Yes	Answer=Deep
Constructive	Novelty=New	Answer=Novel
Impasses	(Doubt=Yes	(Certainness=(Uncertain ∨ Mixed)
	& Accuracy=(Correct ∨ SomeMissing))	& Accuracy=Correct)
	∨ (Accuracy = Incorrect ∨ SomeErrors)	∨ (Accuracy = Incorrect ∨ PartiallyCorr)
Accountable	Depth=Yes	Answer=Deep & Accuracy=Correct
	& (Accuracy=Correct ∨ SomeMissing)	

Finally, to support this mapping of our original annotations[3] and to also ensure the equivalence of all further data analysis, we ported the ITSPOKE data and annotations into the format required by the NITE XML Toolkit (NXT, www.ltg.ed.ac.uk/NITE), which was already used for the BEETLE corpus. NXT comes with a query language and set of command line utilities that cover many of the analyses we require, including the ability to find dialogue contributions matching complex constraints involving tags and their values such as those in Table 1. Our use of NXT, supplemented by common procedures for using NLTK, Excel, SPSS, and a set of unix shell scripts, enabled the use of exactly the same software for all of the data analysis described next.[4]

4. Results: Predicting Learning from Student Dialogue Behaviors

Besides annotation differences, it has also been difficult for the tutorial dialogue community to compare the results of prior studies due to the use of different dependent and independent measures. Our current work always uses posttest score as the dependent measure, and includes pretest score among the independent measures to account for learning gain. A uniform set of additional independent measures for characterizing student dialogue is derived from the annotations described in the prior section.

Our first measures derive from both the word and content annotations. For each student, we count both the total number of student words (**SWords**), as well as the total number of words tagged as content using the same online physics dictionary (**SPhysicsDictWords**), across all dialogues with a student. For the BEETLE corpus only, we also compute content using two alternative methods. First, we use the BEETLE-specific glossary rather than the common physics dictionary to tag domain content words (**SBeetleGlossWords**). Second, we use the manually annotated segments of contentful talk to tag student words as content words (**SManualContentWords**). Finally, we normalize all raw counts by dividing by the total number of (student and tutor) **Words**, across all dialogues with a student.

Table 2 shows the results of using partial correlations to examine the relationship between posttest and each of these independent measures, after first regressing out the correlation with pretest. Correlations that are statistically significant ($p < .05$) are highlighted in **bold**. The first row shows that in BEETLE, students on average contributed 44% of the words when conversing with a (human) tutor. Furthermore, this percentage showed a trend ($p < .1$) to positively correlate with learning (R=.34, p=.08). In IT-

[3] Inter-coder reliability was evaluated for the original annotations, but not yet for the mapped constructs.

[4] However, the output of the automatic tokenizer was hand-corrected for BEETLE, but not for ITSPOKE.

Table 2. Partial Correlations with Posttest (controlled for Pretest)

Measure	**BEETLE** (n=30)			**ITSPOKE** (n=20)		
	Mean	R	p	Mean	R	p
SWords/Words	.44	.34	.08	.04	.17	.48
SPhysicsDictWords/Words (not stemmed)	.05	.22	.26	**.01**	**.60**	**.01**
SBeetleGlossWords/Words (not stemmed)	**.20**	**.38**	**.04**	NA	NA	NA
SManualContentWords/Words	**.38**	**.43**	**.02**	NA	NA	NA
SPhysicsDictWords/Words (stemmed)	.09	.25	.20	.02	.28	.24
SBeetleGlossWords/Words (stemmed)	**.21**	**.38**	**.04**	NA	NA	NA

SPOKE, students contributed only 4% of the words in the dialogues with the ITSPOKE (computer) tutor. Although there is also a positive correlation with learning, it is weaker and not significant at even a trend level. Note that the proportion of student talk is much lower in ITSPOKE than in BEETLE, which we believe reflects the use of computer rather than human tutoring,[5] as well as other differences (spoken versus typed, mechanics versus electricity and other curriculum differences, student population, etc.).

Despite these differences, when only student *content* words are counted, there are significant positive correlations with learning in both corpora. The second row shows that when the same online dictionary is used, the correlation is only significant for ITSPOKE (R=.6, p=.01). This is not surprising given that the dictionary is designed for the "World of Physics". The next two rows show that with a more appropriate dictionary the correlation is also significant in BEETLE (R=.38, p=.04), and becomes even stronger when content words are derived via manual rather than automated methods (R=.43, p=.02). The use of the BEETLE-specific glossary increases the percentage of content words from 5% to 20%, while the use of the manual annotations almost doubles this percentage. Finally, the last two rows show that while adding a stemming preprocessor also increases the percentage of student content words in both corpora, the significant correlation with learning does not improve in BEETLE, and even loses significance in ITSPOKE.

Our next set of measures derive from the dimensions-to-construct mappings shown in Table 1. We first count the number of student contributions that match each of the four learning theory constructs (**Effort, Constructive, Impasses, Accountable**), across all dialogues with a student. We also count the number of contributions that match certain value combinations for the original dimensions; although similar to our original studies, we now achieve more uniformity across corpora by using the tag value correspondences identified by our construct mappings. For example, a new accuracy measure **Right** (student contributions that are correct even if incomplete) counts contributions that match the accuracy values specified in the second conjunct in the **Accountable** mapping in Table 1, for each original tagset. Finally, as with the word and content measures, we normalize each of the raw counts to create percentage measures, e.g. **%Effort**.

Table 3 shows the results of using multivariate linear regression to predict posttest score (the dependent measure) from pretest and these two sets of additional independent measures (constructs and dimensions). The two models shown, and all included parameters (determined automatically using stepwise regression in SPSS), are significant at $p < .05$. The first row shows the included parameters when the set of independent measures includes pretest and the percentage of student contributions matching each of the four

[5] When students interacted with a human tutor, they contributed over 21% of the words in the dialogues [4].

Table 3. Regressions with Posttest

Measures	**BEETLE** (n=30)			**ITSPOKE** (n=20)		
	Predictors	R^2	p	Predictors	R^2	p
Pretest, Constructs	%Impasses (-)	.22	.01	%Accountable (+), %Effort (-)	.50	.01
Pretest, Tags	%Right (+)	.46	.00	%Right (+)	.23	.03

learning constructs. In BEETLE, the best model contains only the single predictor **%Impasses**, which is negatively predictive of learning (R^2=.22, p=.01). In ITSPOKE, a combination of two different parameters best predicts learning (R^2=.50, p <.01). Interestingly, while accountable talk (**%Accountable**) positively predicts learning, student effort (**%Effort**) is negatively correlated. The second row shows the automatically included parameters when the independent measures are now pretest and percentage of student contributions matching specific values of the four DeMAND dimensions (e.g., defining **Right** as **Accuracy**=(*Correct*∨*SomeMissing*) in Beetle). The same model is learned for both corpora: the more accurate the student (**%Right**), the higher the posttest score.

While the results in the first row suggest that different constructs best predict learning for BEETLE and ITSPOKE (at least using the stepwise procedure), we were interested in knowing whether the predictors identified for one corpus would nonetheless still be predictive of learning in the other corpus. To examine this, we use the stepwise parameter selection mechanism to select predictors providing the best model fit for one corpus, then test the utility of this selection by fitting a new linear model containing only these selected parameters to the other corpus. Performing a regression on the BEETLE data using only the parameters selected for ITSPOKE (**%Accountable** and **%Effort**) yields a model that is marginally significant overall (R^2=.18, p < .07), with both predictors (**%Accountable** (+) and **%Effort** (-)) remaining individually significant (p <.03).

5. Discussion and Future Work

Our study has increased our understanding of whether and how prior research findings generalize across multiple tutorial dialogue corpora. We first presented methods for uniformly annotating and statistically analyzing two previously collected corpora in conceptual physics domains: the BEETLE typed human tutoring corpus for electricity and electronics, and the ITSPOKE spoken computer tutoring corpus for qualitative mechanics. We then reexamined our previous findings regarding student dialogue and learning, to see how they might change once the measures and statistical methods were merged.

Replicating our original findings, we again find that in both corpora, the more contentful a student's dialogue contributions are, the more students learn. In addition, our results suggest that simple natural language processing techniques can be used to automatically annotate content, with little performance degradation compared to the use of manual annotation, at least in BEETLE. Our ITSPOKE and BEETLE comparisons, in turn, highlight the importance of using appropriate domain dictionaries.

We also find that with similar coding and analysis, several new results emerge across corpora. While originally student accuracy positively correlates with learning only in BEETLE, once we define the same correctness measure (including both correct utterances and partially correct utterances with no errors), this correlation now holds in ITSPOKE as well. Finally, a new ITSPOKE analysis in terms of learning constructs shows

that both accountable talk and student effort are together predictive of learning; furthermore, the predictive utility of this parameter selection generalizes to BEETLE.

We are currently examining whether our prior results regarding tutor behaviors (e.g., questioning [5] and restating [6]) also generalize across our corpora, and plan to extend our automatic content tagging to use more sophisticated natural processing techniques.

Acknowledgments

This research was supported by a Visiting Professorship Award from the Leverhulme Trust. Thanks to G. Campbell, N. Steinhuaser, and C. Callaway for useful conversations.

References

[1] M. G. Core, J. D. Moore, and C. Zinn. The role of initiative in tutorial dialogue. In *Proc. Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2003.

[2] S. Katz, D. Allbritton, and J. Connelly. Going beyond the problem given: How human tutors use post-solution discussions to support transfer. *International Journal of Artificial Intelligence and Education*, 13, 2003.

[3] S. Craig, A. Graesser, J. Sullins, and B. Gholson. Affect and learning: an exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29(3):241–250, 2004.

[4] D. Litman, C. Rosé, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman. Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education*, 16:145–170, 2006.

[5] D. J. Litman and K. Forbes-Riley. Correlations between dialogue acts and learning in spoken tutoring dialogues. *Journal of Natural Language Engineering*, 12(2):161–176, 2006.

[6] M. Dzikovska, G. Campbell, C. Callaway, N. Steinhauser, E. Farrow, J. Moore, L. Butler, and C. Matheson. Diagnosing natural language answers to support adaptive tutoring. In *Proc. International FLAIRS Conference*, 2008.

[7] K. VanLehn, P. W. Jordan, C. Rosé, D. Bhembe, M. Böttner, A. Gaydos, M. Makatchev, U. Pappuswamy, M. Ringenberg, A. Roque, S. Siler, R. Srivastava, and R. Wilson. The architecture of Why2-Atlas: A coach for qualitative physics essay writing. In *Proc. International Conference on Intelligent Tutoring Systems (ITS)*, 2002.

[8] G. E. Campbell, J. D. Moore, and M. Dzikovska. The ETUDE project, 2007. Presentation at the ONR Cognitive Science PI Meeting.

[9] A. Purandare and D. Litman. Content-learning correlations in spoken tutoring dialogs at word, turn and discourse levels. In *Proc. International FLAIRS Conference*, 2008.

[10] G. E. Campbell, N. B. Steinhauser, M. Dzikovska, J. D. Moore, C. B. Callaway, and E. Farrow. The "DeMAND" coding scheme: A "common language" for representing and analyzing student discourse. In *Proc. International Conference on Artificial Intelligence in Education (AIED)*, 2009.

[11] K. VanLehn, S. Siler, C. Murray, T. Yamauchi, and W. B. Baggett. Why do only some events cause learning during human tutoring? *Cognition and Instruction*, 21(3), 2003.

[12] K. Forbes-Riley, M. Rotaru, and D. Litman. The relative impact of student affect on performance models in a spoken dialogue tutori ng system. *User Modeling and User-Adapted Interaction*, 2008.

[13] S. B. Nolen. Reasons for studying: Motivational orientations and study strategies. *Cognition and Instruction*, 1988.

[14] T. M. Duffy and D. H. Jonassen. *Constructivism and the technology of instruction: A conversation*. Lawrence Erlbaum Associates, 1992.

[15] M. K. Wolf, A. C. Crosson, and L. B. Resnick. Accountable talk in reading comprehension instuction. Technical report, University of Pittsburgh, 2006.

[16] K. Forbes-Riley, D. Litman, and M. Rotaru. Responding to student uncertainty during computer tutoring: A preliminary evaluation. In *Proc. International Conference on Intelligent Tutoring Systems (ITS)*, 2008.