A Study of Analogy in Computer Science Tutorial Dialogues

Mehrdad Alizadeh[1], Barbara Di Eugenio[1], Rachel Harsley[1], Nick Green[1],
Davide Fossati[2]and Omar AlZoubi[2]
[1]*Computer Science Department, University of Illinois at Chicago, Chicago, USA*
[2]*Computer Science, Carnegie Mellon University in Qatar, Doha, Qatar*
{*maliza2, bdieugen, rharsl2, ngreen21*}*@uic.edu*, {*dfossati, oalzoubi*}*@cmu.edu*

Keywords: Computer Science Tutoring, Tutorial Dialogues, Analogy, Linear Regression Analysis, Data Structures.

Abstract: Analogy plays an important role in learning, but its role in teaching Computer Science has hardly been explored. We annotated and analyzed analogy in a corpus of tutoring dialogues on Computer Science data structures. Via linear regression analysis, we established that the presence of analogy and of specific dialogue acts within analogy episodes correlate with learning. We have integrated our findings in our ChiQat-Tutor system, and are currently evaluating the effect of analogy within the system.

1 INTRODUCTION

Learning by analogy is a mechanism by which features of a known concept are mapped to features of an unknown concept (Gentner, 1998). Analogy plays a major role in learning, to the point that some researchers consider it the core of cognition (Hofstadter, 2001). It can be effective in the early stages of learning especially when learners may lack appropriate prior knowledge (Gentner et al., 2003). Our interest in analogy concerns its role in one-on-one tutoring: first, as a strategy employed by human tutors, and second, as a mechanism to be used in Intelligent Tutoring Systems (ITSs). Our specific domain is introductory Computer Science (CS).

As other researchers, we explore human tutoring interactions among tutor and student (Fox, 1993; Chi et al., 2001) with two goals. From the cognitive point of view, we wish to understand how learning is supported by specific strategies the human tutor uses; from a technological point of view, we explore how those successful strategies can be modeled and/or approximated computationally. As concerns analogy, (Nokes and VanLehn, 2008) showed that providing prompts during analogical comparison improve students performance. (Gadgil and Nokes, 2009) applied analogical comparison of worked-out examples and showed that analogy supports collaborative learning especially where conceptual understanding is essential. As far as we know, few ITSs employ analogy, and none within CS. The Bridging Analogies tutor (Murray et al., 1990) utilizes intuitive physical scenarios to explain less intuitive concepts. (Lulis et al., 2004) described implementation of analogies in an ITS for Cardiovascular Physiology. (Chang, 2014) proposed an instructional comparison approach using an analogy tutor that can help conceptual learning for procedural problem solving.

We are developing ChiQat-Tutor, a novel ITS in the domain of basic CS data structures (Fossati, 2013). ChiQat-Tutor has modules on recursion, linked lists, and binary search trees. It provides us with an environment in which we can experiment with different tutoring strategies that we have studied in a corpus of 54 human tutoring sessions on introductory CS data structures, collected in the late 2000's. Strategies we have experimented with in the past include different types of feedback and worked-out examples (Chen et al., 2011; Di Eugenio et al., 2013; Fossati et al., 2015). In order to assess whether ChiQat-Tutor should employ analogies, we investigated learning from analogy in our corpus, as described in Sections 2 (the corpus) and 3 (examples of analogy and annotation thereof). Section 4 presents our results concerning whether analogy is a useful learning strategy for CS data structures.

2 CORPUS

Our corpus contains 54 one-on-one tutoring dialogues on basic computer science (CS) data structures such as stacks, linked lists and binary search trees. Each individual student participated in only one tutoring session, for a total of 54 students participating. The concepts were tutored by one of two tutors, LOW and JAC (in our examples, ST is the student). LOW was an expert tutor in CS with about 30 years of teaching experience while JAC was a senior undergraduate student in CS. A tutoring session took about 37.6 minutes on average. Most of the time the tutor talks producing 93.5% of the total words. Right before each session, the student's prior knowledge was tested via a pre-test. Then, the score for each topic was normalized to the [0..1] interval. The tutor was given a general description of the student's performance on each data structure. In other words, the tutor was not given the numeric scores but was provided with qualitative information to inform his tutoring. Table 1 shows the number of sessions (N) a topic was tutored for. As can be seen, the tutors chose to tutor on lists and trees in almost every session, whereas they sometimes decided to skip stacks. The length of a session is shown in terms of mean (μ) and standard deviation (σ). As can be seen, stacks take the minimum amount of time on average with a mean of 5.8 minutes, whereas lists and BSTs take 14.4 and 19.2 minutes respectively. Please refer to (Chen et al., 2011) for more information about our corpus.

Table 1: Number of sessions and length by topic

Topic	N	Length (min)	
		μ	σ
Lists	52	14.4	5.8
Stacks	46	5.8	1.8
Trees	53	19.2	6.6
All	54	37.6	6.1

3 ANALOGY

Analogy is defined as drawing similarities between different subjects (Gentner, 1998; Gentner and Colhoun, 2010). A similarity usually forms a relation between a known subject and an unknown one that may result in further inferences about the unknown subject. In our tutoring dialogues, both tutors use a set repertoire of analogies. For example, to explain stacks, JAC uses Lego as analogy (Figure 1), whereas LOW uses a stack of trays (Figure 2). For lists, JAC explains the concept by demonstrating the way people stand in a line (Figure 3). The example shows how JAC, during a tutoring session, keeps talking about analogy with a line for a large number of utterances. As concerns trees, both tutors employ family trees as analogy (Figure 4).

3.1 Annotation

In order to study analogy, two annotators annotated our corpus. The annotators were instructed to annotate for the beginning and the end of an analogy episode within a session. For every topic in a session a tutor usually uses at most one analogy such as analogy of Legos for stacks or analogy of line for lists. However, the tutor may refer to the analogy several times during a tutoring session. Consequently, in a session, annotators may annotate several episodes for a specific analogy. We used the Kappa statistics κ (Carletta, 1996; Di Eugenio and Glass, 2004) to measure the level of agreement between annotators. Because our annotators code for analogy episodes, our κ computation is based on the number of lines that annotators annotate similarly or differently. Annotators first double coded 15 sessions five sessions at a time; every five sessions, they revised a predefined manual. After these 15 sessions, they annotated another five sessions. Intercoder agreement was computed on these five sessions, and an acceptable level was reached ($\kappa = 0.58$). κ is negatively affected by skewed data, which occurs because analogy episodes comprise a very small subset of a session. The remainder of the corpus was independently annotated by each annotator (half each).

This annotation task was very challenging due to the difficulty of analogy boundary detection. In fact, annotators are in full agreement on whether analogy is used for a data structure in a session, or not; disagreements are just on the boundaries of analogy episodes.[1]

3.2 Distributional statistics

Table 2 shows basic statistics on analogy for each topic such as number of analogies ($N_{Analogy}$), percentage of sessions an analogy was used in, and length in words of analogy per session. Stacks are the data structure for which analogy is most frequently used. Analogy is used less for lists and the least for trees. However, analogy used for lists is the longest: this is due to the line analogy used by JAC, who refers back to it during the session. As concerns trees being the least likely to be tutored by analogy, we speculate it may be due to the fact that the technical terminology used for trees (parent, daughter, etc), is derived from family trees to start with.

Figure 5 illustrates when in the session the analogy occurs for each topic. As can be seen, the tutor

[1] Disagreement on boundaries of episodes is a well known problem in discourse analysis (Passonneau and Litman, 1997).

219 JAC think of the stack as a bunch of Legos, okay?
220 JAC and each time you put out a Lego...
221 JAC okay, we'll call this, we'll just go a, b, c, d, e, and so forth.
222 JAC okay?
223 JAC so we're stacking our Legos up.
224 JAC if we want to take a Lego off we can only take the Lego off that we just inserted.
225 JAC right?
226 JAC because we're building from the bottom up.
227 JAC okay?
228 JAC so we can only, take in, take off whatever we put in last.

Figure 1: Analogy with Lego for stacks

587 LOW I think of a bunch of trays in a cafeteria where they have this special really very elegant dispenser the trays are in and all you can see is the trays on top.
588 LOW there's a spring down here, and you take that try off the spring and it' loaded so that the next tray pops up.
589 ST okay.
590 LOW and of course somebody comes out of the washing room and puts new trays in.
591 LOW so the only thing you can do is pop something off instead of trays we're going to have this thing on letters a x b.

Figure 2: Analogy with tray for stacks

Table 2: Analogy basic statistics

Topic	N	$N_{Analogy}$	Percentage	Words/Analogy
Lists	52	21	40%	807.430
Stacks	46	40	87%	185.95
Trees	53	16	30%	73.56

Table 3: Analogy usage by tutor

Topic	JAC	LOW	ALL
Lists	17	4	21
Stacks	18	22	40
Trees	9	7	16

utilizes an analogy mostly at the beginning of a tutoring session; the usage of analogy decreases as the session progresses, especially for stacks.

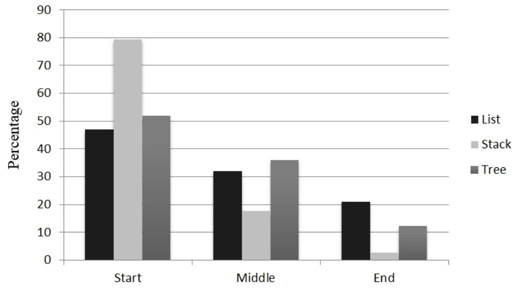

Figure 5: When do analogies occur?

We investigated whether the student's performance in the pre-test may have motivated the tutor to utilize analogy in the tutoring session. Pearson correlation between pre-score and usage of analogy was computed for each topic. Usage of analogy is 1 if analogy was used in a topic during a session otherwise 0. Table 4 shows the results. Negative values for lists and stacks indicate that due to the poor performance of student, the tutor is likely to use analogy.

Table 4: Correlation between analogy and pre-test

Topic	Correlation
Lists	-0.34
Stacks	-0.52
Trees	0.18

4 ANALYSIS

As we noted in the introduction, we are interested in analogy as a tutoring strategy for two reasons: to uncover whether it is effective in this domain, and if so, which of its features are computationally amenable to incorporate in Intelligent Tutoring Systems. Hence, we ran multiple regressions to understand the correlation between analogy and learning gains (following (Litman et al., 2006; Ohlsson et al., 2007)). Our analysis builds on other work we had already done on our data via other dialogue features, specifically, dialogue acts. As the models we will discuss below do indeed include dialogue acts, we first describe the dialogue acts we have coded for in our corpus.

4.1 Dialogue acts

A dialogue act indicates the speaker's intention behind an utterance. There are many reasonable inventories of dialogue acts, including for tutorial dialogues

299	JAC	kind of like a line right?
300	ST	uh huh.
301	JAC	I know who's in front of me but I don't know who's behind me.
.		
433	JAC	um if you need to insert, if you need to switch places with something you use a temporary um you always need to know who's in front of you and then you're going to insert something tell whoever's being inserted what's in front of you and then point to the one you're being inserted to.

Figure 3: Analogy with line for lists

022	JAC	so if we think about it as a family tree um this could be the great grand this would be the great grandparent of this child here.

Figure 4: Analogy with family tree for trees

DDI	TUT	Now a binary search tree must remain ordered.
DPI	TUT	say we want to insert, um, six.
SI	ST	down there? [pointing to tree drawing]
PF	TUT	right
SI	ST	five is smaller than six
DDI	TUT	and the right child of five is null
DPI	TUT	so we will insert six to its right

Figure 6: Example of annotation for dialogue acts

(Litman et al., 2006; Ezen-Can and Boyer, 2013). As discussed in (Chen et al., 2011), our corpus was examined for impressions and trends. Then, given the directive style of our tutors and how much they talked, we defined a minimal set of dialogue acts focused on tutors' contributions. Finally, the tutoring corpus was annotated with this set of dialogue acts. Figure 6 shows an example of dialogue act annotation for trees. Each dialogue act is described below:

- *Positive Feedback (+FB)*: The tutor confirms that the student has performed a correct step.
- *Negative Feedback (-FB)*: The tutor helps the student recognize and correct an error.
- *Direct procedural instruction (DPI)*: The tutor directly informs the student what steps to perform.
- *Direct declarative instruction (DDI)*: The tutor provides facts about the domain or a particular problem.
- *Prompt (PT)*: The tutor attempts to engender a meaningful contribution from the student.
- *Student Initiative (SI)*: The student takes the initiative, namely, produces a dialogue contribution which is not in answer to a tutor's question or prompt.

4.2 Results

Learning gains are computed based on the normalized difference between each student's scores in post-test and pre-test. Multiple linear regression models let researchers explore specific aspects of tutors' behaviors and tutoring interaction properties. Since these models help distinguish effective from non-effective features, they can provide guidelines for the design and development of effective ITSs.

Table 5: Description of features used in multiple linear regression models

Feature	**Description**
Pre-test	Pre-test score
Length	Length of the tutoring session
FB	Number of positive/negative feedbacks
DPI	Number of direct procedural instructions
DDI	Number of direct declaration instructions
PT	Number of prompts
SI	Number of student initiations
AN	If analogy used 1, otherwise 0
AN-Length	Length of analogy episodes in words
AN-FB	Number of positive/negative feedbacks inside analogy episodes
AN-DPI	Number of direct procedural instructions inside analogy episodes
AN-DDI	Number of direct declaration instructions inside analogy episodes
AN-PT	Number of prompts inside analogy episodes
AN-SI	Number of student initiations inside analogy episodes

Table 5 describes the features used in this work, whereas Table 6 includes the statistically significant models we obtained by running all possible combinations of features. β is the correlation coefficient between each feature and learning gain, p shows the level of significance and R^2 explains how well features can describe variations in learning gain. For each topic, Model i represents a statistically significant multiple linear regression model that selects a set of features resulting in the highest adjusted R^2. Model 1 only includes the pre-test, since it is well-known that previous knowledge is a reliable predictor of post-test performance. Pre-test score always has the highest contribution in the best models, indicating that it explains a sizable amount of variance in learning. The correlation is negative which is due to the ceiling effect: students with higher previous knowledge have less chance to learn.

Model 2 includes the pre-test, length of the dialogue and dialogue acts. For lists, positive feedback (PF) correlates with learning; this replicates previous results of ours (Chen et al., 2011; Fossati et al., 2015), and in fact ChiQat-Tutor already includes PF. Additionally, direct procedural instruction (DPI) has a significantly positive correlation with learning. In stacks, prompt (PT) and DPI have positive correlation with learning. PT indicates the importance of the tutor's role in asking students to contribute. In trees, just DPI is statistically significant. DPI contributes in all topics indicating the importance of explaining to students steps to perform a task.

Model 3 includes the pre-test, length of the dialogue, dialogue acts and existence of analogy (AN). Interestingly, AN appears in all of the most predictive models for each topic. AN positively correlates with learning gain, supporting the hypothesis that the usage of analogy helps learning. However, considering the *p* value, it is not statistically significant for stacks and trees. Interestingly, dialogue acts similar to those in Model 2, contribute in Model 3. Model 4 adds analogy-based features resulting in improved adjusted R^2. In lists, student initiative inside analogy correlates with learning. This shows that the students' proactive contributions during an analogy episode have a positive effect on learning. In stacks, DPI inside analogy correlate with learning, confirming step-by-step analogy-based instruction. In trees, PT and direct declarative instruction (DDI) inside analogy correlate with learning. This may suggest that for trees a simple description of analogy with family tree would lead to learning.

Based on the analysis on human-human tutoring, we found that analogy is an effective strategy for learning these three data structures. Hence, we implemented analogy in ChiQat-Tutor, as we briefly describe below.

Table 6: Multiple linear regression models

Topic	Model	Predictor	β	R^2	*p*
List	1	Pre-test	-0.466	0.202	$< .001$
	2	Pre-test	-0.466	0.353	$< .001$
		PF	0.011		$< .1$
		DPI	0.003		$< .1$
	3	Pre-test	-0.330	0.388	$< .001$
		Length	0.011		ns
		PF	0.015		$< .1$
		DPI	0.005		$< .1$
		PT	-0.003		ns
		AN	0.145		$< .1$
	4	Pre-test	-0.41	0.472	$< .01$
		PF	0.007		ns
		DPI	0.29		$< .01$
		DDI	-0.001		ns
		AN	0.298		$< .01$
		AN-Length	0.002		ns
		AN-PF	0.002		ns
		AN-PT	-0.025		ns
		AN-SI	-0.070		$< .1$
Stack	1	Pre-test	-0.462	0.296	$< .001$
	2	Pre-test	-0.495	0.332	0
		PT	0.010		$< .1$
		DPI	0.007		$< .1$
	3	Pre-test	-0.408	0.348	$< .001$
		DPI	-0.01		$< .1$
		PT	0.007		$< .1$
		AN	0.137		ns
	4	Pre-test	-0.443	0.459	$< .001$
		PF	0.025		ns
		DDI	-0.001		ns
		PT	0.025		ns
		SI	0.004		ns
		AN-DPI	0.031		$< .1$
		AN-PF	0.035		ns
Tree	1	Pre-test	-0.742	0.677	$< .001$
	2	Pre-test	-0.703	0.689	$< .001$
		DPI	-0.002		$< .001$
		PT	0.001		ns
	3	Pre-test	-0.755	0.696	$< .001$
		Length	-0.004		ns
		DDI	0.001		ns
		AN	0.069		ns
	4	Pre-test	-0.756	0.729	$< .001$
		AN-Length	-0.006		ns
		AN-DDI	0.038		$< .1$
		AN-PT	-0.117		$< .1$

5 CONCLUSIONS AND CURRENT WORK

In this work, we annotated and analyzed analogy in a corpus of tutoring dialogues on Computer Science data structures. Two annotators annotated analogy episodes (inter-coder reliability was $\kappa = 0.58$). One finding was that the frequency of analogy differed among topics: significantly more analogies were used in stacks and lists than in trees. Additionally, analogies occurred more frequently at the beginning of the dialogue in stacks than in lists and trees. Finally, we used regression analysis to explore whether analogy correlates with learning. Usage of analogy in a session correlates with students' learning gains for lists. Furthermore, some dialogue acts (DAs) that occur within analogy episodes results in more explanatory models that correlate with learning.

Given these results, we have integrated our findings within the ChiQat-Tutor system, for linked lists.

We have integrated analogy in the system in two fashions. First approach is based on the fact that analogy is most likely to appear at the beginning of a session. Similarly when a student starts working with the linked list tutorial, the system displays a window describing an analogy of people standing in a line. Second approach is based on the fact that the tutor frequently refers to analogy during a tutoring session. Correspondingly, for every problem a step by step analogy based example was fashioned that student could refer to. We have recently run a controlled experiment with three conditions: The First condition provides an analogy at the beginning of the session. The Second condition enables student access to analogy based step by step examples for every problem. The Third condition enables student access to worked out examples for every problem. We are currently analyzing the results of these experiments to uncover whether our implementation of analogy is effective.

ACKNOWLEDGEMENTS

This work is supported by award NPRS 5-939-1-155 from the Qatar National Research Fund.

REFERENCES

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

Chang, M. D. (2014). Analogy tutor: A tutoring system for promoting conceptual learning via comparison. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Chen, L., Di Eugenio, B., Fossati, D., Ohlsson, S., and Cosejo, D. (2011). Exploring effective dialogue act sequences in one-on-one computer science tutoring dialogues. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 65–75. Association for Computational Linguistics.

Chi, M. T., Siler, S. A., Jeong, H., Yamauchi, T., and Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science*, 25(4):471–533.

Di Eugenio, B., Chen, L., Green, N., Fossati, D., and Al-Zoubi, O. (2013). Worked out examples in computer science tutoring. In *Artificial Intelligence in Education*, pages 852–855. Springer.

Di Eugenio, B. and Glass, M. (2004). The kappa statistic: A second look. *Computational linguistics*, 30(1):95–101.

Ezen-Can, A. and Boyer, K. E. (2013). In-context evaluation of unsupervised dialogue act models for tutorial dialogue. In *Proceedings of SIGDIAL*, pages 324–328.

Fossati, D. (2013). Chiqat: An intelligent tutoring system for learning computer science. In *Qatar Foundation Annual Research Conference*, number 2013.

Fossati, D., Di Eugenio, B., Ohlsson, S., and Brown, C. (2015). Data driven automatic feedback generation in the ilist intelligent tutoring system. *Technology, Instruction, Cognition and Learning*, (To appear).

Fox, B. A. (1993). *The Human Tutorial Dialogue Project: Issues in the design of instructional systems*. Lawrence Erlbaum Associates Hillsdale, NJ.

Gadgil, S. and Nokes, T. (2009). Analogical scaffolding in collaborative learning. In *annual meeting of the Cognitive Science Society, Amsterdam, The Netherlands*.

Gentner, D. (1998). Analogy. *A companion to cognitive science*, pages 107–113.

Gentner, D. and Colhoun, J. (2010). Analogical processes in human thinking and learning. In *Towards a theory of thinking*, pages 35–48. Springer.

Gentner, D., Loewenstein, J., and Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology*, 95(2):393.

Hofstadter, D. R. (2001). Analogy as the core of cognition. *The analogical mind: Perspectives from cognitive science*, pages 499–538.

Litman, D. J., Rosé, C. P., Forbes-Riley, K., VanLehn, K., Bhembe, D., and Silliman, S. (2006). Spoken versus typed human and computer dialogue tutoring. *International Journal of Artificial Intelligence in Education*, 16(2):145–170.

Lulis, E., Evens, M., and Michael, J. (2004). Implementing analogies in an electronic tutoring system. In *Intelligent Tutoring Systems*, pages 751–761. Springer.

Murray, T., Schultz, K., Brown, D., and Clement, J. (1990). An analogy-based computer tutor for remediating physics misconceptions. *Interactive Learning Environments*, 1(2):79–101.

Nokes, T. J. and VanLehn, K. (2008). Bridging principles and examples through analogy and explanation. In *Proceedings of the 8th international conference on International conference for the learning sciences-Volume 3*, pages 100–102. International Society of the Learning Sciences.

Ohlsson, S., Di Eugenio, B., Chow, B., Fossati, D., Lu, X., and Kershaw, T. C. (2007). Beyond the code-and-count analysis of tutoring dialogues. *Artificial intelligence in education: Building technology rich learning contexts that work*, 158:349.

Passonneau, R. J. and Litman, D. J. (1997). Discourse segmentation by human and automated means. *Computational Linguistics*, 23(1):103–139.