From Strangers to Partners: Examining Convergence within a Longitudinal Study of Task-Oriented Dialogue

Christopher M. Mitchell **Kristy Elizabeth Boyer** **James C. Lester**

Department of Computer Science
North Carolina State University
Raleigh, NC, USA
{cmmitch2, keboyer, lester}@ncsu.edu

Abstract

Convergence is thought to be an important phenomenon in dialogue through which interlocutors adapt to each other. Yet, its mechanisms and relationship to dialogue outcomes are not fully understood. This paper explores convergence in textual task-oriented dialogue during a longitudinal study. The results suggest that over time, convergence between interlocutors increases with successive dialogues. Additionally, for the tutorial dialogue domain at hand, convergence metrics were found to be significant predictors of dialogue outcomes such as learning, mental effort, and emotional states including frustration, boredom, and confusion. The results suggest ways in which dialogue systems may leverage convergence to enhance their interactions with users.

1 Introduction

Convergence is a widely observed phenomenon in dialogue, in which interlocutors adapt to the patterns in each other's utterances (Brennan 1996; Pickering and Garrod 2004). These patterns can include lexical choice (Hirschberg 2008; Ward and Litman 2007), syntactic choice (Reitter et al. 2006; Stoyanchev and Stent 2009) and loudness (Coulston et al. 2002). It is believed that convergence is indicative of shared understanding (Pickering and Garrod 2004), which makes it an important consideration for task-oriented dialogue systems.

In addition to facilitating shared understanding, convergence has also been associated with the success of dialogues in several domains (Steinhauser et al. 2011; Ward and Litman 2007), and can also be leveraged for lexical and syntactic priming that may improve performance of spoken dialogue systems via more accurate speech recognition (Stoyanchev and Stent 2009). While such results have established that convergence is an important dialogue phenomenon, the field does not yet fully understand how convergence is associated with dialogue success.

This paper examines surface-level and lexical convergence within textual task-oriented dialogues. The analysis considers three levels of convergence: utterance-level *short-term* priming effects, *conversation-level* convergence effects, and *longitudinal* convergence effects, as interlocutors participate in six conversations together over the course of several weeks. Using these measures, we build multiple regression models that indicate ways in which convergence can predict both desirable and undesirable outcomes of task-oriented dialogues.

This paper makes several contributions. First, by examining convergence at several granularity levels and across multiple dialogues with the same partners, we gain insight into how convergence phenomena unfold over time. Second, the findings provide confirmatory evidence that in some domains, such as the tutorial dialogue considered here, lexical priming be associated with unintended consequences. Finally, we demonstrate that dialogue convergence is also associated with affective components such as frustration, engagement, and confusion. These results contribute to an understanding of convergence that may enable us to harness this phenomenon more effectively within dialogue systems.

2 Related Work

Convergence and the related concepts of alignment and priming have been extensively studied. Alignment, or the development of shared understanding, has been studied by Pickering and Garrod (2004) who propose that alignment on lower-level observable features is indicative of alignment at the level of conceptual models. The influence of shared representation in dialogue has also been explored in the context of learning; for example, Ward and Litman (2007) studied lexical convergence in human-human tutoring and found that the rate of priming, which measures student re-use of tutor words at various distances, was positively associated with learning for students with low initial test scores. Conversely, Steinhauser et al. (2011) analyzed lexical convergence in an automated dialogue-based physics tutor, and found that the level of the student mimicking the tutor was negatively correlated with learning. Thus, the relationship between dialogue convergence and learning is not fully understood, and may be highly dependent on context.

In addition to a theoretical link to shared representations, convergence has practical implications, in particular for speech recognition (Stoyanchev and Stent 2009). Brennan (1996) found that users adapt their lexical choices to match those of an automated system in both text-based and speech-based interactions, even when it is apparent that the system understood the user's original lexical choice. Convergence has even been found to occur in non-lexical aspects of a dialogue, such as users adapting their loudness levels to match that of a software agent (Coulston et al. 2002). Together, these results suggest that convergence has implications beyond lexical and syntactic choice.

3 Corpus

The corpus consists of text-based tutorial dialogues between two interlocutors, a tutor and a student, working together to complete tasks in the domain of introductory computer science (excerpt in Appendix A). The corpus was collected over two semesters, in which 67 first-year university students were selected from an introductory engineering course and assigned to one of seven tutors of varying levels of tutoring experience. Each student engaged in six task-based dialogues with a single tutor over four weeks with the goal of producing a working software artifact during each session. Each session included several subtasks, and time was strictly limited to forty minutes duration. The remote collaboration interface, shown in Figure 1, facilitated a real-time synchronized view of the workspace and dialogue. This paper considers dialogue utterances only, leaving to future work the analysis of task-related artifacts.

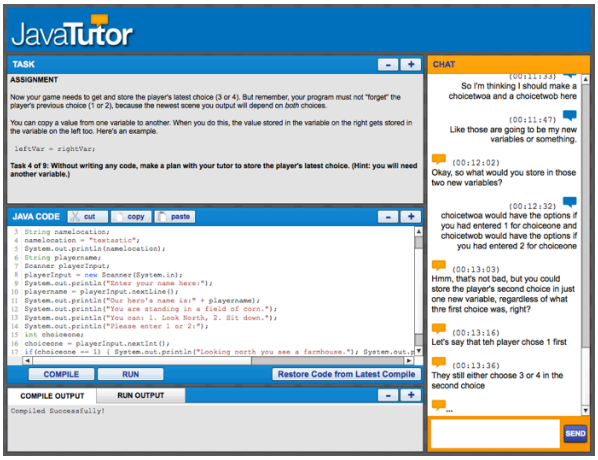

Figure 1. Task-oriented dialogue interface

The effectiveness of the dialogue was measured in several ways. First, student learning was measured as difference in score on pre-test and post-tests. Student engagement, or level of involvement during the dialogue, was measured with a brief survey after each dialogue (O'Brien and Toms 2010), as were student's satisfaction with the exchange, and a rating of how mentally challenging the task was perceived to be (Hart and Staveland 1988). Finally, the tutors were asked to rate their satisfaction with the effectiveness of each session and to report on their perceptions of the affective states of both interlocutors during the session. The students were not asked about their own affective states, as this may have introduced bias in subsequent dialogues.

4 Analysis

The goal of the analysis is to identify the characteristics of the dialogues that are predictive of the outcomes of interest, including learning, engagement, affect, and overall success of the

dialogue as rated by the interlocutors. Summary statistics for the dialogues were computed, including time duration of the session, number of utterances, number of words, number of characters, mean word length, and lexicon size (Table 1). Stop words were not excluded from the analysis, in part due to specialized usage of common vocabulary in the computer science domain (e.g., *for*, *if*).

Although not traditionally considered a form of convergence, we were interested in the relationship between the levels of activity of the two interlocutors. To this end, we analyzed the number of utterances, words, and characters used by tutor and student, and found a significant positive correlation on these metrics ($p<0.0001$ for each).

The first convergence phenomenon considered centers on lexical priming, the tendency for one interlocutor to re-use words previously introduced by the other. We have utilized a priming metric computed as follows: Interlocutor A's *Priming Ratio* (PR) is the percent of Interlocutor A's words reused by Interlocutor B at a given distance *d*, where distance is measured in terms of number of Interlocutor B's utterances. Negative slope of PR over distance indicates a priming effect because an interlocutor was more likely to reuse a word shortly after its use by the other interlocutor. This metric has been used to investigate tutor priming (Steinhauser et al. 2011; Ward and Litman 2007), and we generalize it to measure priming for both interlocutors. Note that student PR, which reflects the extent to which the tutor adopted the student's lexical choice, is of particular interest from the perspective of dialogue system design, in which tutor utterances are system-generated.

	Tutor mean (SD)	**Student mean (SD)**
Surface Features		
Number of utterances	83.7 (28.8)	35.6 (13.1)
Number of words	580.9 (202.3)	170.1 (92.6)
Number of characters	2383.4 (886.6)	667.3 (386.0)
Mean word length	4.1 (0.2)	3.9 (0.3)
Lexicon size	329.7 (87.3)	106.3 (47.3)
Convergence Metrics		
Priming Ratio (1-10)	.030 (.02)	.047 (.02)
ΔPriming Ratio (1-10)	-.011 (.02)	-.017 (.04)
Max Priming Ratio	.052 (.02)	.091 (.04)
Matched Word Ratio	.233 (.09)	.386 (.08)

Table 1. Statistics for each metric

In addition to the Priming Ratio, we also computed a metric to reflect convergence: Interlocutor A's *Matched Word Ratio* (MWR) is the percent of Interlocutor A's words that had been previously used by Interlocutor B at any point in the dialogue history. Because it is backward-looking, this metric is applicable not only in a corpus study, but could also be used within a runtime system to track convergence as the dialogue unfolds.

5 Models and Results

Mean Matched Word Ratio for both interlocutors increased as sessions progressed, reflecting that the two dialogue partners used more of each other's words as they spent more time together. The Priming Ratio also revealed several phenomena in the corpus. Similarly to prior observations from tutorial dialogue (Ward and Litman 2007), we found that student reuse of tutor primes decreased with distance, indicating that a lexical priming effect occurred (Figure 2). This trend also occurred for tutor reuse of student primes (Figure 3). The effect was more pronounced in the tutor's PR than the student's PR; that is, there was more evidence that tutors converged to students in the short term. This finding may be associated in part with the higher number of tutor utterances: a distance in terms of number of tutor utterances represents fewer combined student and tutor utterances than the same distance in terms of student utterances. Additionally, tutor convergence may reflect a dimension of intentional pedagogical choice.

The Priming Ratio is designed to reflect short-term priming. However, there is evidence of a longer-term effect as the two interlocutors engaged in dialogue across multiple sessions. Figures 2 and 3 display Tutor's PR and Student's PR, respectively, by task set, of which there were six in the corpus study. The last task set displays an overall higher level of lexical convergence than the earlier sessions, and there is a general trend of increasing convergence as the number of sessions together increases.

In order to identify the features that were most predictive of dialogue outcomes, all of the convergence metrics and surface summary features were provided as input to a stepwise linear regression model. Standard greedy variable addition and removal was performed, with additional post-processing and re-training to eliminate instances of multicollinearity. The learned models (Appendix B) include a mixture of

convergence metrics and surface features, as well as structural features such as the task set number and the time duration of the dialogue. At least one convergence metric was found to be associated with each outcome in the generated models, with the exception of Engagement.

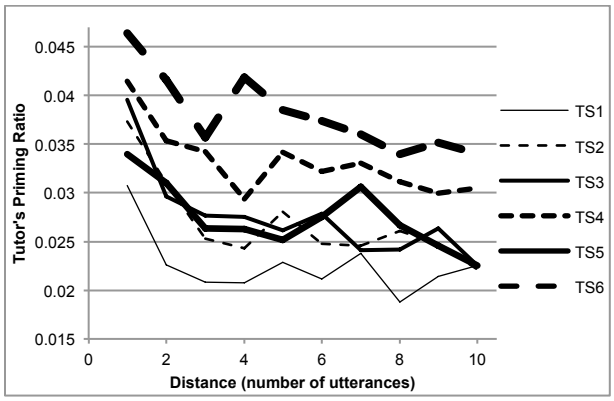

Figure 2. Tutor's Priming Ratio aggregated by task set (TS = Task Set)

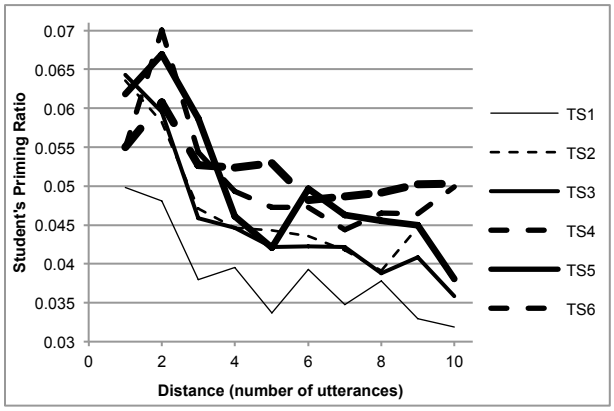

Figure 3. Student's Priming Ratio aggregated by task set (TS = Task Set)

Several significant relationships emerged within the models. We discuss a subset of these here. First, tutor Priming Ratio was a significant predictor for outcomes as rated by both tutor and student. Higher tutor Priming Ratio was associated with higher tutor perception of dialogue success, perhaps because students reflected tutor lexical choice more frequently. The same metric was associated with lower student score for how mentally demanding the tasks were perceived to be, which suggests that a shared lexicon may be associated with decreased cognitive load.

Another significant finding is the relationship between student Priming Ratio and student boredom, confusion, and frustration. In all the models, increased reports of these student emotions by the tutor corresponded to lower student Priming Ratio. This result suggests that tutor reuse of student lexical choice may be associated with positive affective outcomes.

Finally, the tutor's Matched Word Ratio is a significant negative predictor of learning gains, and also a significant negative predictor for student confusion. This finding may be related to the fact that by reusing more student language, the tutor may be effectively introducing fewer novel contributions that might lead to confusion.

6 Conclusion and Future Work

Understanding how convergence unfolds holds significant promise for designing more effective dialogue systems. Toward that end, this paper has explored convergence in task-oriented dialogue at three levels: at the level of pairs of utterances, across a single conversation, and over multiple conversations with the same interlocutors. The results demonstrate that within the corpus, the two interlocutors display increasing levels of convergence longitudinally. Additionally, the results suggest ways in which short-term and long-term convergence are associated with particular positive and negative aspects of dialogue success and user affect.

The findings have significant implications for dialogue systems. First, they suggest that not only may successful lexical priming aid in understanding (Stoyanchev and Stent 2009), it may also be associated with lower cognitive load for users. Additionally, it may be possible to leverage convergence to positively impact users' affective states with respect to emotions such as boredom, confusion, and frustration. These potential relationships suggest that work to further elucidate convergence phenomena is particularly promising because dialogue systems stand to benefit from strategically leveraging convergence and adaptation.

Acknowledgments

This work is supported in part by the National Science Foundation through Grants DRL-1007962 and CNS-1042468. Any opinions, findings, conclusions, or recommendations expressed in this report are those of the participants, and do not necessarily represent the official views, opinions, or policy of the National Science Foundation.

References

Brennan, S. (1996). Lexical Entrainment in Spontaneous Dialog. In *Proceedings of the 1996 International Symposium on Spoken Dialogue*, 41-44.

Coulston, R., Oviatt, S., and Darves, C. (2002). Amplitude Convergence in Children's Conversational Speech with Animated Personas. In *Proceedings of the 7th International Conference on Spoken Language Processing*, 2689–2692.

Hart, S. and Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P.A. Hancock and N. Meshkati, eds., *Human Mental Workload*. 1988, 139-183.

Hirschberg, J. (2008). High Frequency Word Entrainment in Spoken Dialogue. In *Proceedings of ACL HLT*, 169-172.

O'Brien, H. and Toms, E. (2010). The Development and Evaluation of a Survey to Measure User Engagement. *Journal of the American Society for Information Science and Technology*, 6(1), 50-69.

Pickering, M. and Garrod, S. (2004). Toward a Mechanistic Psychology of Dialogue. *Behavioral and Brain Sciences*, 27(2), 169-226.

Reitter, D., Moore, J., and Keller, F. (2006). Priming of Syntactic Rules in Task-Oriented Dialogue and Spontaneous Conversation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 685-690.

Steinhauser, N., Campbell, G., Taylor, L., Scott, C., Dzikovska, M., and Moore, J. (2011). Talk Like an Electrician: Student Dialogue Mimicking Behavior in an Intelligent Tutoring System. In *Proceedings of the 15th International Conference on Artificial Intelligence in Education*, 361-368.

Stoyanchev, S. and Stent, A. (2009). Lexical and Syntactic Priming and Their Impact in Deployed Spoken Dialog Systems. In *Proceedings of NAACL HLT*, 189-192.

Ward, A. and Litman, D. (2007). Dialog Convergence and Learning. In *Proceedings of the 13th International Conference on Artificial Intelligence in Education*, 262-269.

Appendix A. Corpus Excerpt

T: yes so what happens with the other paths?

S: is it because the last statement is fullfilled so it has no need to print the error?
S: i understand what is happening but i do not know how to explain it
T: ok so you noticed that when the if statement directly before it is true then it does not go to the else
T: but if the if statement directly before the else statement is false then it goes to the else statement
S: yes.

S: so i need to make all of them else if statements?

T: yes

Appendix B. Regression Models

	β	*p*
Norm. Learning Gain, R^2 = .0687		
Tutor's MWR	-.169	.0160
Task set number	-.144	.0392
Engagement (Student-reported) R^2 = .0892		
Tutor's number of characters	-.527	.0007
Student's mean word length	.159	.0033
Tutor's mean word length	-.169	.0053
Tutor's lexicon size	.369	.0189
Mentally demanding (Student-reported) R^2 = .217		
Tutor's PR (distances 1-5)	-.128	.0118
Session length (ms)	.174	.0065
Combined number of utterances	.579	.0005
Tutor's number of utterances	-.475	.0040
Tutor's number of characters	-.439	.0031
Tutor's mean word length	-.118	.0496
Tutor's lexicon size	.627	<.0001
Student confusion*, R^2 = .319		
Student's PR (distances 1-10)	-.233	<.0001
Tutor's number of matched words	1.04	<.0001
Tutor's MWR	-.523	<.0001
Task set number	-.122	.0105
Session length (ms)	.292	<.0001
Student's number of characters	.247	.0048
Combined lexicon size	-.594	<.0001
Student frustration*, R^2 = .300		
Max value of Student's PR	.156	.0035
Session length (ms)	.239	<.0001
Tutor's number of utterances	.460	<.0001
Tutor's number of words	.342	.0135
Tutor's lexicon size	-.748	<.0001
Student boredom*, R^2 = .202		
Student's PR (distances 1-5)	-.234	<.0001
Tutor's number of utterances	.261	.0001
Tutor's lexicon size	-.412	<.0001
Session successful overall*, R^2 = .246		
Tutor's PR (distances 1-3)	.186	.0002
Δ Student's PR (distances 1-10)	.122	.0079
Session length (ms)	-.420	<.0001
Tutor's number of utterances	.518	<.0001
Tutor's number of words	-.473	.0006
Tutor's lexicon size	.275	.0340

* = from tutor perception survey;
β = standardized regression coefficient