Automatically Measuring Lexical and Acoustic/Prosodic Convergence in Tutorial Dialog Corpora

Arthur Ward and Diane Litman

Learning Research and Development Center
University of Pittsburgh
Pittsburgh, Pa., 15260, USA
artward@cs.pitt.edu, litman@cs.pitt.edu

Abstract

We use language technology to develop corpus measures of lexical and acoustic/prosodic convergence. We show that these measures successfully discriminate randomized from naturally ordered data, and demonstrate both lexical and acoustic/prosodic convergence in our corpus of human/human tutoring dialogs.

1. Introduction

Human users of computer dialog systems have been shown to exhibit a wide variety of speech convergence behaviors. In this work we use "convergence" as a general term for the tendency of dialog partners to adjust various features of their speech to be more similar to one another. This phenomenon is of general interest to researchers in dialog systems for several reasons. For example, it may be possible to enlist convergence to improve speech recognition rates [1, 2], or increase user compliance with system requests [3]. Convergence also may be of particular interest to researchers in tutorial dialog systems because of the predictions of the Interactive Alignment Model (IAM) [4]. This model suggests that observable convergence is the result of interactive priming between dialog partners. The theory also suggests that convergence at observable lexical and acoustic/prosodic levels may accompany alignment at the higher semantic levels. We hypothesize that if users converge toward the productions of a tutorial dialog system, their convergence may be associated with learning.

There is experimental evidence that human users do indeed converge toward (non tutorial) dialog systems lexically [5], as well as in acoustic/prosodic features of speech such as amplitude [1]. These studies measure convergence by comparing groups of users in different experimental conditions. To train dialog systems, however, we need corpus measures of convergence. In this paper we describe two new corpus measures of amplitude and lexical convergence which we adapt from previous work in the literature. We show that students converge toward their human tutor in these aspects of speech. In separate work [6], we show that these measures are useful predictors of learning. The measures of convergence we use are adapted from work by Reitter, Keller and Moore [7].

Reitter et al. [7] demonstrate a method for measuring the effects of syntactic priming in dialog corpora. "Priming" refers to the mechanism the IAM holds responsible for the convergence of various speech properties. Hearing and decoding a speech unit, such as a certain word or syntactic structure for example, "primes," (ie: increases the activation of) corresponding internal representations. If these representations are still active during the next speech production, they are more likely to be used than alternatives which are less active. The speech unit that caused the increase in activation is called a "prime." In this work we refer to the increased usages following the prime as the "response." Reitter et al. measure the response as an increased use of primed syntactic structures. We first adapt their method to measure response as the increased use of primed words. We then further adapt it to measure acoustic/prosodic changes following unusually extreme tutor utterances.

2. The Corpus

2.1. Ordered Data

Our data is taken from a corpus of human-human tutoring transcripts collected by the ITSPOKE intelligent tutoring system group for a previous study [8]. In these tutoring sessions, a human tutor presents a problem in qualitative physics to a student, who answers it in essay form. The tutor examines this essay, identifies flaws in it, and engages the student in a tutorial dialog to remediate those flaws. A single tutor helped fourteen students do up to ten problems each. This resulted in a corpus of 128 dialogs, which contain 6,721 student turns and 5,710 tutor turns. In this work we look for student responses to primes in tutor utterances, and we ignore essay turns.

2.2. Random Data

We also created a corpus of randomized tutoring dialogs for use as a baseline. This corpus contains the same set of tutoring dialogs as our ordered corpus. In each dialog, tutor utterances are left in their original position, but the positions of all student utterances are randomized within the dialog. Because the student utterances are no longer in their original relationship to the tutor utterances, we expect to find reduced priming effects in this corpus. A successful measure of priming effects should give positive results on the ordered data, but not on this randomized data.

3. Lexical Convergence

3.1. Lexical Measure

For our measure of lexical convergence, we count any word uttered by the tutor as a potential prime. Following Reitter et al. [7], we define the next N student turns as a window in which to look for a student response. However, where Reitter counted repetition of syntactic rules as the response measure, we count the student's use of the tutor's prime word. If the tutor's prime word occurs

	Speaker	Transcript	Student Response to Prime	
			Words	Data Points
1	Tutor	seat is in contact seat exerts a force so the result is that torso is accelerating in forward direction now what will happen to the head?		
2	Student	oh, ok		
3	Student	**the head will** move back because it's not attached **to the seat**	the, head, will, to, seat	[2,6]

Table 1: Portion of a transcript, student response to tutor primes, and data points generated

Data	Window Size			
	5	10	15	20
Ord.	-.026 (<.26)	**-.042** (<.0001)	**-.027**(<.0001)	**-.022** (<.0001)
Rand.	-.002 (.91)	.001 (.88)	.002 (.67)	.002 (.63)

Table 2: Slopes and (p-values) for lexical measures, counting all tokens. P-values and slopes below adjusted threshold in bold

once in the first utterance of the student's response window, for example, we count a response of one at distance one. This process is illustrated in Table 1. Our lexical measure would first take the tutor utterance shown in row 1 as the prime, set the next N student turns to be the response window, and count the number of prime repetitions in each turn of that window. For example, the second student utterance after this prime contains six repetitions of prime words. This is at a distance of two from the prime, and so generates the data point [2,6].

The student response window following each tutor utterance in the corpus is examined in this way, generating a set of data. We then use linear regression to determine the relationship between distance from the prime and lexical repetition count. Linear regression produces the slope of a fitted line and a p-value that indicates the probability of fitting that line if there were really no relationship between distance and response.

In this work we make thirty-six comparisons, looking for evidence of convergence. Therefore, we apply the Bonferroni correction to reduce the chance of a type one error. We will consider p-values below .0014 (.05/36) to be significant. P-values below this threshold are shown bold in all tables.

Results for our lexical measure are shown in Table 2. The top row gives the slope of the fitted line for the ordered data. Each slope is followed by its p-value in parentheses. Columns 2 - 5 give results for each of four response window sizes. The bottom row of Table 2 gives results on the randomized corpus, in the same format. We selected window sizes starting at 20, toward the top of the range used by Reitter et. al. [7], and working downward until the fitted slopes became non-significant. This produced four windows for lexical results, however "what is an optimal window size?" and "why?" are still open research questions. The slopes are all negative for the ordered lexical data, indicating convergence immediately after the prime, which then decays with distance.

Our randomized data show no significant results. Our metric is sensitive to utterance sequence, which suggests that the convergence we find in the ordered corpus is real.

3.2. Priming and Lexical Convergence

The IAM describes several levels at which dialog partners may align, with alignment at one influencing alignment at neighboring levels. Therefore, when we measure lexical convergence it is not clear to what extent we are measuring the effects of lexical priming or of alignment at semantic or other levels. Measures of convergence may be more useful if we can determine the type of priming involved. So, as a first step toward identifying the effects of lexical priming, we attempt to remove the effect of words for which there had been no other choice. Ideally, this task would involve doing perfect word-sense disambiguation, but in this work we use the following, simpler approach. Each word is marked with its part of speech and all synsets for the word in that POS are retrieved from WordNet [9]. To identify words for which there may have been no alternative choices, we count the number of synonyms in each synset. If no synset contains more than one choice, we consider it probable that there was no suitable alternative word available to the speaker, and remove that response from the data. This adjustment reduced the number of data points collected by 47%, from 25,352 to 13,415. 33,387 tokens were skipped in the corpus, representing 240 different word types. The majority of tokens identified this way were particles and other closed set words not included in WordNet. The left two columns of Table 3 show the nine most frequent words in this set. Together, they account for almost 75% of the tokens skipped.

Removing these words makes intuitive sense: in the student utterance in Table 1, the student's productions of "the" may have been made necessary by the use of "seat" and "head," rather than as the result of independent priming. This measure does, however, fail to identify other probable instances of "no-choice" words because it lacks word sense disambiguation. That is, a word may have no alternatives in the intended sense, but this measure may find alternatives in another sense, and so fail to remove the word. This adjustment should therefore be considered a first approximation, which probably has high precision but lower recall.

Results for our lexical measure, skipping these "no-choice" words, are shown in the top half of Table 4. Because we have removed one source of our measured convergence, the significant slopes become slightly more shallow. Slopes for randomized data still give non-significant p-values, though these p-values are much smaller than before the correction.

Some instances of lexical repetition in our dialogs may also be a topic effect. That is, regardless of any lexical priming effects, the students may have tended to repeat certain words simply because they were talking about the same subject as the tutor. Therefore, we next attempt to further isolate the effects of lexical priming by also removing the effect of topic. To do this, we combine two lists of "physics specific" words collected for previous projects. The first list includes physics topic titles culled from a publicly available physics web site [10][1]. The second list was collected for previous work [11]. Combined, these lists contain 1,085 physics-related terms. For our "topic" correction, we do not count student

[1]We thank Amruta Purandare for her generosity in compiling this list.

Words Skipped				Words Counted	
No-choice		Topic			
word	#	word	#	word	#
the	9664	force	1133	on	2622
it	3222	velocity	831	some	1332
is	2440	acceleration	754	job	973
uh	2236	horizontal	384	saying	896
that	1930	time	383	word	839
you	1699	motion	379	about	828
to	1304	direction	330	become	652
of	1226	equal	271	rise	520
and	1040	law	254	he	463

Table 3: Lexical priming corrections: Top 75% of words skipped for lack of choice, top 27% of words skipped for topic correction, top 68% of words counted after these adjustments

repetitions of tutor words if they appear on this list.

Results for our lexical measure, skipping both "no-choice" and "topic" words, are shown in the bottom half of Table 4. Removing this source of convergence makes the fitted slopes more shallow, although still significant on the ordered data. P-values on randomized data remain above our corrected significance threshold.

The center two columns of Table 3 show the top 27% of additional words skipped under the "topic" correction. Again, this correction makes intuitive sense, many of these words seem to be terms made necessary by the physics topic under discussion.

Even after making these two corrections, however, substantial lexical choice remains in the corpus. For example, after the first correction, lexical variability is visible among non-physics terms. In the following utterance a student uses both "greater" and "larger" to indicate an increased extent.

> "so, that'll cause the acceleration to be greater and the, um, wait let me think for a second, um, the acceleration will be larger in the, in the small, in the lightweight, has a less mass"

And after the second correction, students show a variety of other words for physics terms such as "accelerate:"

> "it will pick up won't it pick up speed?"

The convergence we measure after these two corrections seems to represent the temporary reduction in this lexical variety, which may result in part from lexical priming. Finally, we present the repeated student words which remain after these two corrections. The right two columns of Table 3 show the nine most frequent words *remaining* after this adjustment, which account for 68% of the tokens counted.

4. Acoustic/Prosodic Convergence

To generate data from which to measure acoustic/prosodic (**AP**) convergence, we calculated RMS amplitude (loudness) and f0 (pitch) values for each tutor and student turn in our corpus. For both RMS and f0 we calculated the max, min, and mean value over each turn. Neither measure was normalized, partly because normalization of these features had not been helpful in previous work [12]. Mean RMS was also used in the convergence study of Coulston et al. [1]. F0 is interesting in part because it is also

Data	Window Size			
	5	10	15	20
No-choice Correction				
Ord.	-.030 (.03)	**-.025** (<.0001)	**-.018** (<.0001)	**-.012** (<.0001)
Rand.	-.014 (.32)	.000 (.96)	-.004 (.12)	-.005 (.012)
No-choice & Topic Correction				
Ord.	-.013 (.23)	**-.015** (<.0001)	**-.010** (<.0001)	**-.007** (<.0001)
Rand.	-.011 (.40)	-.002 (.69)	-.005 (.05)	-.004 (.022)

Table 4: Slopes and (p-values) for lexical measures: Top:"no choice" words removed. Bottom: "no-choice" and "topic" words removed. P-values and slopes below adjusted threshold in bold

	Window Size			
Max RMS	**15**	**20**	**25**	**30**
Ord. slope:	-7.2884	-16.1007	**-19.6304**	**-16.1840**
Ord. pVal:	0.4284	0.0091	**<.0001**	**<.0001**
Rand. pVal:	0.3080	0.3755	0.7052	0.1495
Mean RMS	**15**	**20**	**25**	**30**
Ord. slope:	-3.3547	-4.1780	-4.8891	**-4.0174**
Ord. pVal:	0.2764	0.0449	0.0016	**0.0010**
Rand. pVal:	0.7081	0.6242	0.7307	0.5178
Min RMS	**15**	**20**	**25**	**30**
Ord. slope:	0.4830	0.4077	0.2153	0.1963
Ord. pVal:	0.2460	0.1518	0.3161	0.2553
Rand. pVal:	0.6557	0.4753	0.4827	0.7224

Table 5: Results for RMS amplitude measures. P-values and slopes below adjusted threshold in bold.

automatically computable, and might be available in a future affect recognizing tutor [13]. For acoustic/prosodic convergence, we used the same window size selection procedure as for the lexical measure, but also added two larger windows because the starting window size of 20 student turns was already too short to find significant slopes in the RMS data.

As described in section 3.1, our lexical measure used word repetition as the response variable by counting up the number of repeating words in each utterance in the student response window. We now use the same approach for measuring responses in acoustic/prosodic data. Instead of recording the number of repeated words at each distance d from the prime, however, we record the value of the acoustic/prosodic variable at each distance.

We cannot use the same definition of a prime as we did in lexical data, however. Now, instead of having a discrete trigger like word occurrence, we have continuous acoustic/prosodic values. To identify a "prime" in this data, we turn to Fisher's Z score, a standardized measure of distance from the mean often used to detect outliers. Z is calculated as $\frac{x-\mu}{\sigma}$ where x is the acoustic/prosodic value for the current turn, μ is the population mean, and σ is the population standard deviation. We locate a prime wherever the tutor's AP value had a Z score greater than one, meaning it was more than one standard deviation above its mean. This threshold setting counted 486 tutor utterances as "primes" using the maxRMS feature. Using a threshold reflects the intuition that we want to measure the student's response to unusually loud tutor utterances. Finding the exact threshold setting that produces the most useful measure, however, is a topic for future research.

Min f0	Window Size			
	15	20	25	30
Ord. slope:	**0.4594**	**0.2800**	**0.2544**	**0.2650**
Ord. pVal:	**0.0002**	**0.0012**	**<.0005**	**<.0005**
Rand. pVal:	0.3047	0.1758	0.1983	0.9242

Table 6: Results for f0 (pitch) measures. P-values and slopes below adjusted threshold in bold.

Results for the loudness (RMS) features, with the prime set at Z>1, are shown in Table 5. Table 5 is divided horizontally to give results for our three RMS features, maximum RMS, mean RMS and minimum RMS. Within each of those three divisions are three rows. The top row shows the slope of the fitted line on ordered data. The second row shows the p-value of that slope. The third line shows the p-values of lines fitted to our randomized data.

Results for max RMS are similar to those obtained on lexical data. We see that window sizes larger than 20 give significant, negative slopes on ordered data. Mean RMS becomes significant at a window size of 30, also with a negative slope. Neither of these measures produce significant results on the randomized data. For the min RMS feature we have no significant results.

Table 6 shows results for minimum f0, locating primes where $Z > 1$. Here the pattern of results is different. Neither the max nor mean f0 features gave significant results, and are not shown. The min f0 feature, on the other hand, produced significant results in ordered data for all window sizes, but with a positive slope. It gave no significant results on randomized data.

5. Discussion and Future Work

We have proposed two new measures of convergence based on one developed by Reitter et al. [7] to detect syntactic priming. We first extended their measure to detect lexical convergence, and introduced further modifications to help isolate the effects of lexical priming. We showed lexical convergence both before and after these "no-choice" and "topic" adjustments. We next extended this measure to detect acoustic/prosodic convergence by using a threshold to identify primes. We showed evidence for convergence of max and mean RMS. Evidence for the success of these measures comes from the negative slope of their fitted regression lines on ordered data, from the significant p-values of those lines, and from their lack of false-positive results on randomized data.

Our "no-choice" and "topic" adjustments were no doubt only partially successful in isolating the effects of lexical priming. We hope to improve these measures by, for instance, removing not only "no-choice" words but also a set of "most-frequent" words taken from a large corpus. However, while separating the effects of the various levels of priming is interesting from a theoretical perspective, it may not be necessary for the measures described to be useful in dialog systems. We have shown in separate work, [6] for example, that several of the measures described here are useful predictors of learning in tutorial dialog. In particular the slope of the lexical response line with a window size of 20 was a useful predictor of learning in two separate corpora of tutoring dialogs, one with a computer and the other with a human tutor. Also, the slope of the mean RMS response line was found to predict learning for students with high pre-test scores, for both human and computer tutors. We are currently extending this work to include different tutors, in different tutoring domains. Following that, we hope to make the ITSPOKE tutor aware of student convergence behavior and able to adjust instruction based on the amount of convergence measured.

6. Acknowledgments

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

7. References

[1] 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.

[2] L. Bell, J. Gustafson, and M. Heldner, "Prosodic adaptation in human-computer interaction," in *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS 03)*, Barcelona, Spain, 2003.

[3] D. Buller and R. Aune, "The effects of speech rate similarity on compliance: Application of communication accommodation theory," *Western Journal of Communication*, vol. 56, pp. 37–53, 1992.

[4] M. J. Pickering and S. Garrod, "Toward a mechanistic psychology of dialogue," in *Behavioral and Brain Sciences*, vol. 27, 2004.

[5] S. Brennan, "Lexical entrainment in spontaneous dialog," in *Int. Symposium on Spoken Dialog*, 1996, pp. 41–44.

[6] A. Ward and D. Litman, "Dialog convergence and learning," in *Proceedings 13th International Conference on Artificial Intelligence Education (AIED)*, Los Angeles, Ca., 2007.

[7] D. Reitter, F. Keller, and J. Moore, "Computational modelling of structural priming in dialogue," in *Proceedings of the Human Language Technology Conference of the NAACL, companion volume*, 2006, pp. 121–124.

[8] D. J. Litman, C. P. Rosé, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman, "Spoken versus typed human and computer dialogue tutoring," in *Proceedings of the 7th International Conference on Intelligent Tutoring Systems(ITS)*. Maceio, Brazil, 2004.

[9] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International Journal of Lexicography (special issue)*, vol. 3 (4), pp. 235–312, 1990.

[10] Eric W. Weisstein, "Eric weisstein's world of physics. http://scienceworld.wolfram.com/physics," 2006.

[11] A. Ward and D. Litman, "Predicting learning in tutoring with the landscape model of memory," in *Proceedings of the 2nd Workshop on Building Educational Applications using NLP*, Ann Arbor, June 2005, pp. 21–24.

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

[13] D. J. Litman and K. Forbes-Riley, "Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors," in *Speech Communication*, vol. 48 (5), May 2006, pp. 559–590.