Measuring Convergence and Priming in Tutorial Dialog

Arthur Ward **Diane Litman**
Learning Research and Development Center
University of Pittsburgh
Pittsburgh, Pa., 15260, USA
artward@cs.pitt.edu litman@cs.pitt.edu

Abstract

Experimental research has shown that human users will converge with dialog systems along many dimensions of speech, including those of acoustic/prosodic features and lexical choice. Other results suggest that speech convergence may provide a variety of benefits to spoken dialog systems, such as an improved user model, increased ease of use, improved feelings of intimacy, and increased compliance on the part of the user. These potential benefits to dialog systems of generating or detecting convergence behaviors suggest the need for corpus studies of convergence, in addition to the experimental results. Here, we build on previous work to demonstrate corpus measures of lexical and acoustic/prosodic convergence. We show that these measures successfully distinguish randomized from naturally ordered data, and demonstrate both lexical and acoustic/prosodic priming effects 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 is of general interest to researchers in dialog systems for several reasons. Convergence of word choice, produced by lexical entrainment, may help explain why humans can produce and disambiguate speech so quickly in conversation (Brennan, 1996). It has also been suggested that the natural convergence of acoustic/prosodic features might be enlisted to lead users to speak in a way that would be more understandable to dialog systems (Coulston et al., 2002; Bell et al., 2003). There is also evidence that perceived speech rate similarity produces feelings of increased immediacy, which are in turn linked to greater compliance with requests for help (Buller and Aune, 1992).

Convergence may also be of particular interest to researchers in *tutorial* dialog systems because of the predictions of the Interactive Alignment Model (Pickering and Garrod, 2004). This model suggests that observable convergence is the result of a process of interactive alignment between dialog partners. It describes the processes of generating and understanding speech as happening on a number of levels, from the lower acoustic/prosodic up to higher semantic and situation model levels. During dialog, conversational partners tend to align their internal representations at each of these levels, with alignment at one level affecting alignment at neighboring levels. Alignment of these internal representations leads to the observable convergence of various speech features. The model therefore suggests that convergence at observable lexical and acoustic/prosodic levels may accompany alignment at the higher semantic levels. From this, we hypothesize that if users converge toward the productions of a speech-enabled tutorial dialog system, their convergence may be associated with learning.

A number of researchers have found experimental evidence that users will in fact converge toward the productions of (non-tutorial) dialog systems in several acoustic/prosodic dimensions of speech. For example, Coulston et al. (2002) found that children would increase their spoken amplitude toward that of an "extroverted" spoken dialog agent, and reduce it toward that of an "introverted" dialog agent. Similarly, children's response latencies have been found to converge toward the text-to-speech productions of a computer partner (Darves and Oviatt, 2002). Users have also been found to adapt their speaking rate to more closely match that of an animated character in a simulated dialog system (Bell et al., 2003).

Convergence has been shown not only in acoustic/prosodic features of speech, but also in lexical choice. Brennan (1996) has found that users are likely to adopt the terms used by a wizard-of-oz dialog system, and that this tendency is at least as strong as with human dialog partners.

The evidence outlined above was generated by well controlled experimental studies. In these studies convergence is detected by comparing the productions of users between experimental conditions. For example, if as in the Coulston et al. study, users speaking to a "loud" system become themselves significantly louder than users speaking to a "soft" system, that is evidence for convergence in amplitude between users and the system. For use in dialog systems, however, we need measures which can detect convergence during dialog. This type of measure can be developed using corpora. In this paper, we describe two new measures which can detect convergence in a corpus of tutoring dialogs. Our measures detect lexical alignment, as studied by Brennan, and also acoustic/prosodic alignment including amplitude, as studied by Coulston et al. We show that students converge toward their human tutor in several of these dimensions. In separate work (Ward and Litman, 2007), we show that these measures are useful predictors of learning.

Our measures are adapted from previous work by Reitter et al. (Reitter et al., 2006b; Reitter et al., 2006a). Reitter et al. demonstrate a method for measuring syntactic priming in dialog corpora. "Priming" refers to the mechanism the Interactive Alignment Model 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" or increases the activation of corresponding internal representations at that level. 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. (2006b) measure priming effects in the following way. They first parse a corpus to associate each utterance with a set of phrase structure rules. Each of these rules is treated as a potential prime. To measure the response to this prime, they count the number of times the same rule recurs in each of the following N utterances. They use logistic regression to model the effect of distance from the prime on the probability of rule repetition. A negative coefficient of distance is interpreted as reflecting an increased probability of rule recurrence immediately after the prime, which decreases back toward the global mean with distance. Using this method, they find corpus evidence for syntactic priming, and find this effect to be stronger in task oriented than in non task-oriented dialog.

We modify Reitter et al.'s method first by counting the repetition of words, rather than of syntactic rules, as the response variable, and use it to measure lexical convergence. We also experiment with two corrections designed to isolate the effects of lexical priming in the observed convergence. Next, we further modify the measure for use with a continuously valued prime, and use it to detect acoustic/prosodic convergence.

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 (Litman et al., 2004). 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. The student then reworks the essay, and the cycle repeats until the tutor is satisfied

	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]
4	Tutor	well it it's not moving back it it		
5	Student	just remains behind him		
6	Tutor	yes it's not there is no **force** acting on it		
7	Student	right		
8	Tutor	so it does not it will not accelerate		
9	Student	right		
10	Tutor	now what **the** person will like **to** do will he like his head to go along with the rest of the body?		
11	Student	yeah		
12	Tutor	ok, so what he should do then?		
13	Student	have **a** headrest **to** apply **the force**	a, to, the, force	[2,2] [4,1] [7,4]

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

that all important points are covered. This procedure was designed to mimic the tutoring sequence used by the tutSys spoken dialog tutoring system (Litman and Silliman, 2004). Fourteen students did up to ten problems each, resulting in a corpus of 128 dialogs. Barge-ins, in which one speaker makes an utterance (often a backchannel acknowledgment) while the other speaker is still talking, were handled by placing the barge-in turn after the interrupted turn, rather than splitting the interrupted turn around the barge-in. An example of this is in turn two of Table 1. This produced a corpus of 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. 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. We present results on both of these data sets, for each of our measures of priming.

3 Lexical Convergence

3.1 Lexical Measure

For our first measure of lexical convergence, we count any word uttered by the tutor as a potential prime. Following Reitter et al. (2006b), 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 once in the second utterance of the student's response window, for example, we count a response of one at distance two. This process is illustrated in Table 1. The left three columns of this table give turn numbers, identify speakers and show their utterances. The fourth column identifies the words counted as student responses to the tutor's prime, which are also shown in bold in column three. If we ran our first lexical measure on this excerpt, it would proceed in the following way. It would first take the tutor utterance shown in row 1 as the prime, and take the next N student turns as the response window. The second student utterance after this prime repeats six of the prime words. This is at a distance of two from the prime, and so generates the data point [2,6]. The next student response to this prime occurs in turn 13, which is at a distance of seven from the prime. The words "a," "the," "to," and "force" repeat in this utterance, generating the data point [7,4], shown in the fifth column. Next we take the tutor utterance in row four to be the prime, and define a new student response window after it. This window contains no responses, so the prime is moved to line six. The word "force" from this prime repeats in student turn 13. This is at a distance of four from the prime, and so generates the data point [4,1]. Next the tutor prime is moved to line eight, which produces no responses, then to line ten. The words "the" and "two" from this prime repeat in the student utterance at line 13. This utterance is at a distance of two from the prime, so this response produces the data point [2,2], as shown in the table.

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 for a fitted line and a p-value which indicates the probability of fitting that line if there were really no relationship between distance and response. Figure 1 shows the results of fitting a line to the data points generated by the example in Table 1. Given those four points, linear regression produces the line $3.417 - 0.0447 * dist$ which has a negative slope.

In this work we fit 36 lines (3 lexical, 3 amplitude and 3 pitch measures, at 4 window sizes each) to our ordered data, 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.

In general, for each measure we examined four window sizes near those reported by Reitter et al. (2006a) and below. On lexical data we started at 25 and worked downward until the fitted lines lost significance. Lines fitted to acoustic/prosodic data lost significance at larger window sizes, so we added a window size of 30 for these measures. Optimal window size is still an open research question. In this study we only examine tutor-to-student priming. Same speaker priming is ignored. So, for example, the tutor's re-use of the word "force" in row six of Table 1 is not counted. Student-to-tutor priming is also left for future work.

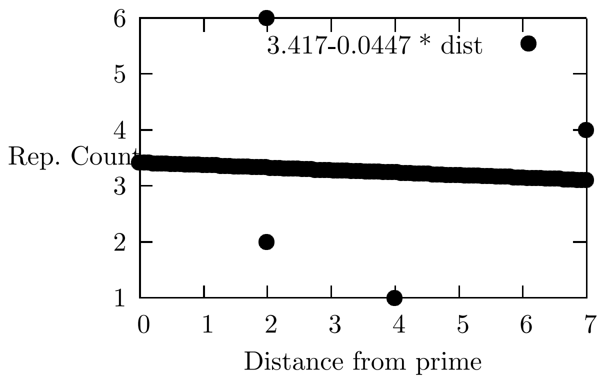

Figure 1: Example regression line fitted to data from table 1

Data	**Window Size**			
Ordered	**5**	**10**	**15**	**20**
slope:	-0.02690	-0.04213	-0.02693	-0.02199
slope pVal:	0.26	**<.0001**	**<.0001**	**<.0001**
Random	**5**	**10**	**15**	**20**
slope:	-0.0026	0.0014	0.0022	0.0016
slope pVal:	0.9122	0.8798	0.6752	0.6352

Table 2: Regression results for lexical measure, counting all primed tokens. P-values below adjusted threshold in bold.

Results for our first lexical measure are shown in Table 2. The top two rows give the slope of the fitted line, and that slope's p value, for the ordered data. Columns 2 - 5 give results for each of four response window sizes. The bottom two rows of Table 2 give results on the randomized corpus, in the same format. We see that a window size of 5 does not give significant results in our ordered corpus, however results for the larger windows are significant. The slopes are all negative for the ordered data, indicating convergence immediately after the prime, which then decays. Interestingly, the three significant slopes shown become less steep with larger window sizes. This suggests that lexical priming effects largely decay within a window size of ten.

On the randomized data, however, there are no significant results. Our metric is sensitive to utterance sequence, which suggests that the priming effects we find in the ordered corpus are real, and not an artifact of data processing.

3.2 Priming and Lexical Convergence

The Interactive Alignment Model describes several levels at which dialog partners may align, with alignment at one level influencing alignment at neighboring levels. Therefore, when we measure lexical convergence it isn't clear if we are measuring the effects of lexical priming, or of alignment at semantic, syntactic, 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 isolating the effects of lexical priming we make two simple adjustments to our measure of lexical convergence. First, we attempt to remove the effect of words for which there had been no other choice. Second, we attempt to remove words which are repeated simply as a "topic effect," because student and tutor are both talking about physics.

We intuitively expect lexical priming to affect word choice, leading us to use primed words when we might otherwise choose an alternative to express the same idea. If so, then using a certain word does not reflect lexical priming (although it may reflect priming at some other level) if there were no other choices in context to express that meaning. As a first step toward isolating the effects of lexical priming, therefore, we attempt to identify words for which there may not have been another choice. We do this in the following way. Each word is tagged for it's part of speech (POS) using a Brill tagger (Liu, 2004), and all synsets for the word in that POS are retrieved from WordNet (Miller et al., 1990). In the event the tagger chooses a POS under which WordNet has no entries, we use all nonempty synsets for that word. To identify words for which there were no alternative choices, we count the number of synonyms in each synset. If no synset contains more than one choice, we consider that there probably was no other word choice available to the speaker, and remove that response from the data.

Results for our lexical measure, skipping these "no-choice" words, are shown in the top half of Table 3. Because we have (at least partially) removed one of the sources of convergence, our measured slopes become more shallow than before. The slopes remain significant in windows larger than 5 on ordered data. P-values on randomized data remain above our Bonferroni adjusted threshold.

	Window Size			
	5	**10**	**15**	**20**
No-choice Correction: Ordered Data				
slope	-0.0305	-0.0252	-0.0180	-0.0128
slope pVal	0.0328	< **0.0001**	< **0.0001**	< **0.0001**
No-choice Correction: Random Data				
slope	-0.0145	0.0003	-0.0045	-0.0050
slope pVal	0.3256	0.9628	0.1283	0.0120
No-choice & Topic Correction: Ordered Data				
slope	-0.01397	-0.01534	-0.01028	-0.00791
slope pVal	0.23071	**0.00016**	< **0.0001**	< **0.0001**
No-choice & Topic Correction: Random Data				
slope	-0.0114	-0.0020	-0.0051	-0.0042
slope pVal	0.4068	0.6928	0.0598	0.0222

Table 3: Lexical conversion corrected for lack of word choice and topic effects. P-values below adjusted threshold in bold.

Removing "no-choice" words makes a substantial difference in the number of data points counted. For example, running our lexical algorithm again, with a window size of 20, but not counting these words reduces 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 4 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 first student utterance in Table 1, for example, 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. However, requiring that a word have no alternatives in any sense is very strict, and often allows words to remain because they have alternatives in senses other than the one intended. For example "K" appears frequently in our transcripts to indicate a shortened, non-inflected utterance of "OK." Our "no-choice" correction does not remove this word because WordNet includes a synset with several synonyms of "K" such as "thousand" and "grand." This adjustment for lexical choice should therefore be considered a first approximation, which probably has high precision but lower recall.

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 (Eric W. Weisstein, 2006)[1]. The second list was collected for previous work (Ward and Litman, 2005). Combined, these lists contain 1,085 physics-related terms. For our "topic" correction, we do not count student 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 3. Removing this additional source of convergence again 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 4 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

[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 4: 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

in this lexical variety, which results from lexical priming. Finally, we present the repeated student words which remain after these two corrections. The right two columns of Table 4 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 estimate acoustic/prosodic (**AP**) convergence, we calculated values for RMS amplitude (loudness) and f0 (pitch) for each tutor and student turn in our corpus. For this we used Entropic Research Laboratory's get_f0 tool, with no post correction. For both RMS and f0 we calculated the max, min and mean over each turn.

As described in section 3.1, our lexical measure counted word repetitions as the response variable. That is, it counted up the number of repeating words at each distance d from the prime, within the student response window. We now use the same approach to 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 an acoustic/prosodic feature at each distance d from the prime.

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 $\frac{x-\mu}{\sigma}$

Max RMS	Window Size 15	20	25	30
Ord. slope:	-7.2884	-16.1007	-19.6304	-16.1840
Ord. pVal:	0.4284	0.0091	**<.0005**	**<.0005**
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 below adjusted threshold in bold.

Max f0	Window Size 15	20	25	30
Ord. slope:	0.1613	0.0589	-0.0525	-0.0052
Ord. pVal:	0.4843	0.7057	0.6528	0.9555
Rand. pVal:	0.6987	0.6616	0.3214	0.8607
Mean f0	**15**	**20**	**25**	**30**
Ord. slope:	0.1053	0.0513	0.0443	0.0777
Ord. pVal:	0.4269	0.5708	0.5120	0.1576
Rand. pVal:	0.3667	0.0619	0.2189	0.2581
Min f0	**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 below adjusted threshold in bold.

where x is the acoustic/prosodic value for the current turn, μ is the population mean, and σ is the population standard deviation. We show results using a threshold of Z = 1, meaning that we locate a prime in any tutor utterance with an a/p value more than one standard deviation above the tutor's mean level for that feature. Using a threshold reflects the intuition that we want to measure the student's response to extreme (either in loudness or pitch) tutor utterances. Finding the exact threshold which gives the most useful measure is a topic for future work.

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. Neither ordered nor randomized data produce significant results for this feature.

Table 6 shows results for the three pitch features, maximum, mean and minimum f0, locating primes where $Z > 1$. Here the pattern of results is different. Neither the max nor mean f0 features gave significant results. 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

We have proposed two new measures of convergence based on one previously developed by Reitter et al. (2006b) to detect syntactic priming. We first extended their measure to detect lexical convergence by counting, as the response variable, the number of times the student adopts the tutor's choice of words. We then made two modifications to this measure intended to isolate the effects of lexical priming. First, we removed the effect of words that had to be used because they had no synonyms. Second, we removed the effect of words which were likely to be used simply because the tutor and student were discussing the same physics topic. Both of these adjustments are first approximations, which could be readily improved, however we believe that they succeed to some extent in isolating the effect of lexical priming.

Finally, we extended this measure again for use with continuously valued acoustic/prosodic data. We introduced a thresholding mechanism to locate primes in those tutor utterances which are unusually loud, or which have unusually extreme pitch features. Using this measure, we showed convergence for max and mean RMS amplitude features. We also found some evidence for *di*vergence in pitch data.

6 Future Work

We have shown in separate work (Ward and Litman, 2007) that some of the measures developed here are useful predictors of learning. We hope to expand on those findings by investigating other features of speech mentioned in the literature, such as spoken tempo or response latency. We also hope to make several improvements to our measures of convergence. In this paper, we explored ways to isolate the effect of lexical priming in observed convergence. We hope to apply this work by isolating and *removing* the effects of various types of non-semantic priming. If this is successful, the remaining convergence may be the effect of semantic alignment, and so be more strongly correlated with learning.

Following that, we hope to use these measures to gather data with which to improve our tutorial dialog system. We expect that dialog system improvements may occur in two broad categories. First, a tutorial system which can detect portions of a dialog with low student convergence may be able to improve learning by offering additional tutoring on those topics.

Second, creating a tutor where speech productions converge toward those of the human user might produce shorter dialogs or more satisfied users. These potential improvements to dialog quality are suggested by the various experimental results mentioned in the introduction. Our current work has looked only at tutor-to-student priming. To investigate possible links between convergence and these measures of dialog quality, we intend also to investigate student-to-tutor priming.

References

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

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

David Buller and R.Kelly Aune. 1992. The effects of speech rate similarity on compliance: Application of communication accommodation theory. *Western Journal of Communication*, 56:37–53.

Rachel Coulston, Sharon Oviatt, and Courtney Darves. 2002. Amplitude convergence in children's conversational speech with animated personas. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP'02)*, Denver, CO.

Courtney Darves and Sharon Oviatt. 2002. Adaptation of users spoken dialogue patterns in a conversational interface. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP'02)*, Denver, CO.

Eric W. Weisstein. 2006. Eric weisstein's world of physics. http://scienceworld.wolfram.com/physics.

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

Diane J. Litman, Carolyn P. Rosé, Kate Forbes-Riley, Kurt VanLehn, Dumisizwe Bhembe, and Scott Silliman. 2004. Spoken versus typed human and computer dialogue tutoring. In *Proceedings of the 7th International Conference on Intelligent Tutoring Systems(ITS)*. Maceio, Brazil.

Hugo Liu. 2004. Montylingua: An end-to-end natural language processor with common sense. web.media.mit.edu/ hugo/montylingua.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to wordnet: An on-line lexical database. *International Journal of Lexicography (special issue)*, 3 (4):235–312.

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

David Reitter, Frank Keller, and Johanna Moore. 2006a. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, companion volume*, pages 121–124.

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

Arthur Ward and Diane Litman. 2005. Predicting learning in tutoring with the landscape model of memory. In *Proceedings of the 2nd Workshop on Building Educational Applications using NLP*, pages 21–24, Ann Arbor, June.

Arthur Ward and Diane Litman. 2007. Dialog convergence and learning. In *Proceedings 13th International Conference on Artificial Intelligence Education (AIED)*, Amsterdam, Netherlands.