Entrainment in Speech Preceding Backchannels

Rivka Levitan
Dept. of Computer Science
Columbia University
New York, NY 10027, USA
rlevitan@cs.columbia.edu

Agustín Gravano
DC-FCEyN & LIS
Universidad de Buenos Aires
Buenos Aires, Argentina
gravano@dc.uba.ar

Julia Hirschberg
Dept. of Computer Science
Columbia University
New York, NY 10027, USA
julia@cs.columbia.edu

Abstract

In conversation, when speech is followed by a *backchannel*, evidence of continued engagement by one's dialogue partner, that speech displays a combination of cues that appear to signal to one's interlocutor that a backchannel is appropriate. We term these cues *backchannel-preceding cues* (BPC)s, and examine the Columbia Games Corpus for evidence of entrainment on such cues. Entrainment, the phenomenon of dialogue partners becoming more similar to each other, is widely believed to be crucial to conversation quality and success. Our results show that speaking partners entrain on BPCs; that is, they tend to use similar sets of BPCs; this similarity increases over the course of a dialogue; and this similarity is associated with measures of dialogue coordination and task success.

1 Introduction

In conversation, dialogue partners often become more similar to each other. This phenomenon, known in the literature as *entrainment*, *alignment*, *accommodation*, or *adaptation* has been found to occur along many acoustic, prosodic, syntactic and lexical dimensions in both human-human interactions (Brennan and Clark, 1996; Coulston et al., 2002; Reitter et al., 2006; Ward and Litman, 2007; Niederhoffer and Pennebaker, 2002; Ward and Mamidipally, 2008; Buder et al., 2010) and human-computer interactions (Brennan, 1996; Bell et al., 2000; Stoyanchev and Stent, 2009; Bell et al., 2003) and has been associated with dialogue success and naturalness (Pickering and Garrod, 2004; Goleman, 2006; Nenkova et al., 2008). That is, interlocutors who entrain achieve better communication. However, the question of how best to measure this phenomenon has not been well established. Most research has examined similarity of behavior over a conversation, or has compared similarity in early and later phases of a conversation; more recent work has proposed new metrics of synchrony and convergence (Edlund et al., 2009) and measures of similarity at a more local level (Heldner et al., 2010).

While a number of dimensions of potential entrainment have been studied in the literature, entrainment in turn-taking behaviors has received little attention. In this paper we examine entrainment in a novel turn-taking dimension: *backchannel-preceding cues* (BPC)s.[1] Backchannels are short segments of speech uttered to signal continued interest and understanding without taking the floor (Schegloff, 1982). In a study of the Columbia Games Corpus, Gravano and Hirschberg (2009; 2011) identify five speech phenomena that are significantly correlated with speech followed by backchannels. However, they also note that individual speakers produced different combinations of these cues and varied the way cues were expressed. In our work, we look for evidence that speaker pairs negotiate the choice of such cues and their realizations in a conversation – that is, they entrain to one another in their choice and production of such cues. We test for evidence both at the global and at the local level.

[1] Prior studies termed cues that precede backchannels, *backchannel-inviting cues*. To avoid suggesting that such cues are a speaker's conscious decision, we adopt a more neutral term.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics:shortpapers, pages 113–117,
Portland, Oregon, June 19-24, 2011.

In Section 2, we describe the Columbia Games Corpus, on which the current analysis was conducted. In Section 3, we present three measures of BPC entrainment. In Section 4, we further show that two of these measures also correlate with dialogue coordination and task success.

2 The Columbia Games Corpus

The Columbia Games Corpus is a collection of 12 spontaneous dyadic conversations elicited from native speakers of Standard American English. 13 people participated in the collection of the corpus. 11 participated in two sessions, each time with a different partner. Subjects were separated by a curtain to ensure that all communication was verbal. They played a series of computer games requiring collaboration in order to achieve a high score.

The corpus consists of 9h 8m of speech. It is orthographically transcribed and annotated for various types of turn-taking behavior, including *smooth switches* (cases in which one speaker completes her turn and another speaker takes the floor), *interruptions* (cases in which one speaker breaks in, leaving the interlocutor's turn incomplete), and backchannels. There are 5641 exchanges in the corpus; of these, approximately 58% are smooth switches, 2% are interruptions, and 11% are backchannels. Other turn types include overlaps and pause interruptions; a full description of the Columbia Games Corpus' annotation for turn-taking behavior can be found in (Gravano and Hirschberg, 2011).

3 Evidence of entrainment

Gravano and Hirschberg (2009; 2011) identify five cues that tend to be present in speech preceding backchannels. These cues, and the features that model them, are listed in Table 1. The likelihood that a segment of speech will be followed by a backchannel increases quadratically with the number of cues present in the speech. However, they note that individual speakers may display different combinations of cues. Furthermore, the realization of a cue may differ from speaker to speaker. We hypothesize that speaker pairs adopt a common set of cues to which each will respond with a backchannel. We look for evidence for this hypothesis using three different measures of entrainment. Two of these measures capture entrainment globally, over the course of an entire dialogue, while the third looks at entrainment on a local level. The unit of analysis we employ for each experiment is an *interpausal unit* (IPU), defined as a pause-free segment of speech from a single speaker, where pause is defined as a silence of 50ms or more from the same speaker. We term consecutive pairs of IPUs from a single speaker *holds*, and contrast hold-preceding IPUs with backchannel-preceding IPUs to isolate cues that are significant in preceding backchannels. That is, when a speaker pauses without giving up the turn, which IPUs are followed by backchannels and which are not? We consider a speaker to use a certain BPC if, for any of the features modeling that cue, the difference between backchannel-preceding IPUs and hold-preceding IPUs is significant (ANOVA, $p < 0.05$).

Cue	**Feature**
Intonation	pitch slope over the IPU-final 200 and 300 ms
Pitch	mean pitch over the final 500 and 1000 ms
Intensity	mean intensity over the final 500 and 1000 ms
Duration	IPU duration in seconds and word count
Voice quality	NHR over the final 500 and 1000 ms

Table 1: Features modeling each of the five cues.

3.1 Entrainment measure 1: Common cues

For our first entrainment metric, we measure the similarity of two speakers' cue sets by simply counting the number of cues that they have in common over the entire conversation. We hypothesize that speaker pairs will use similar sets of cues.

The speakers in our corpus each displayed 0 to 5 of the BPCs described in Table 1 (mean = 2.17). The number of cues speaker pairs had in common ranged from 0 to 4 (out of a maximum of 5). Let S_1 and S_2 be two speakers in a given dialogue, and $n_{1,2}$ the number of BPCs they had in common. Let also $n_{1,*}$ and $n_{*,2}$ be the mean number of cues S_1 and S_2 had in common with all other speakers in the corpus not partnered with them in any session. For all 12 dia-

logues in the corpus, we pair $n_{1,2}$ both with $n_{1,*}$ and with $n_{*,2}$, and run a paired t-test. The results indicate that, on average, the speakers had significantly more cues in common with their interlocutors than with other speakers in the corpus ($t = 2.1$, $df = 23$, $p < 0.05$).

These findings support our hypothesis that speaker pairs negotiate common sets of cues, and suggest that, like other aspects of conversation, speaker variation in use of BPCs is not simply an expression of personal behavior, but is at least partially the result of coordination with a conversational partner.

3.2 Entrainment measure 2: BPC realization

With our second measure, we look for evidence that the speakers' actual values for the cue features are similar: that not only do they alter their production of similar feature sets when preceding a backchannel, they also alter their productions in similar ways.

We measure how similarly two speakers S_1 and S_2 in a conversation realize a BPC as follows: First, we compute the difference ($d^f_{1,2}$) between both speakers for the mean value of a feature f over all backchannel-preceding IPUs. Second, we compute the same difference between each of S_1 and S_2 and the averaged values of all other speakers in the corpus who are not partnered with that speaker in any session ($d^f_{1,*}$ and $d^f_{*,2}$). Finally, if for any feature f modeling a given cue, it holds that $d^f_{1,2} < min(d^f_{1,*}, d^f_{*,2})$, we say that that session exhibits mutual entrainment on that cue.

Eleven out of 12 sessions exhibit mutual entrainment on pitch and intensity, 9 exhibit mutual entrainment on voice quality, 8 on intonation, and 7 on duration. Interestingly, the only session not entraining on intensity is the only session not entraining on pitch, but the relationships between the different types of entrainment is not readily observable.

For each of the 10 features associated with backchannel invitation, we compare the differences between conversational partners ($d^f_{1,2}$) and the averaged differences between each speaker and the other speakers in the corpus ($d^f_{1,*}$ and $d^f_{*,2}$). Paired *t*-tests (Table 2) show that the differences in intensity, pitch and voice quality in backchannel-preceding IPUs are smaller between conversational partners than between speakers and their non-partners in the corpus.

Feature	***t***	***df***	***p*-value**	**Sig.**
Intensity 500	-4.73	23	9.09e-05	*
Intensity 1000	-2.80	23	0.01	*
Pitch 500	-3.38	23	0.002	*
Pitch 1000	-3.28	23	0.003	*
Pitch slope 200	-1.77	23	0.09	.
Pitch slope 300	-0.93	23	N.S.	
Duration	0.50	23	N.S.	
# Words	1.39	23	N.S.	
NHR 500	-2.00	23	0.06	.
NHR 1000	-2.30	23	0.03	*

Table 2: T-tests between partners and their non-partners in the corpus.

The differences between interlocutor and their non-partners in features modeling pitch show that there is no single "optimal" value for a pitch level that precedes a backchannel; this value is coordinated between partners on a pair-by-pair basis. Similarly, while varying intensity or voice quality may be considered a universal cue for a backchannel, the specific values of the production appear to be a matter of coordination between individual speaker pairs.

While some views of entrainment hold that coordination takes place at the very beginning of a dialogue, others hypothesize that coordination continues to improve over the course of the conversation. T-tests for difference of means show that indeed the differences between conversational partners in mean pitch and intensity in the final 1000 milliseconds of backchannel-preceding IPUs are smaller in the second half of the conversation than in the first ($t = 3.44, 2.17$; $df = 23$; $p < 0.05, 0.01$), indicating that entrainment in this dimension is an ongoing process that results in closer alignment after the interlocutors have been speaking for some time.

3.3 Measure 3: Local BPC entrainment

Measures 1 and 2 capture global entrainment and can be used to characterize an entire dialogue with respect to entrainment. We now look for evidence to support the hypothesis that a speaker's realization of BPCs influences how her interlocutor produces BPCs. To capture this, we compile a list of pairs of backchannel-preceding IPUs, in which the second member of each pair follows the first in the conver-

sation and is produced by a different speaker. For each feature, we calculate the Pearson's correlation between acoustic variables extracted from the first element of each pair and the second.

The correlations for mean pitch and intensity are significant ($r = 0.3$, two-sided t-test: $p < 0.05$, in both cases). Other correlations are not significant. These results suggest that entrainment on pitch and intensity at least is a localized phenomenon. Spoken dialogue systems may exploit this information, modifying their output to invite a backchannel similar to the user's own previous backchannel invitation.

4 Correlation with dialogue coordination and task success

Entrainment is widely believed to be crucial to dialogue coordination. In the specific case of BPC entrainment, it seems intuitive that some consensus on BPCs should be integral to the successful coordination of a conversation. Long latencies (periods of silence) before backchannels can be considered a sign of poor coordination, as when a speaker is waiting for an indication that his partner is still attending, and the partner is slow to realize this. Similarly, interruptions signal poor coordination, as when a speaker has not finished what he has to say, but his partner thinks it is her turn to speak. We thus use mean backchannel latency and proportion of interruptions as measures of coordination of whole sessions. We use the combined score of the games the subjects played as a measure of task success. We correlate all three with our two global entrainment scores and report correlation coefficients in Table 3.

Entrain. measure	**Success/coord. measure**	***r***	***p*-value**
1	Latency	-0.33	0.06
	Interruptions	-0.50	0.01
	Score	0.22	N.S.
2	Latency	-0.61	0.002
	Interruptions	-0.22	N.S.
	Score	0.72	6.9e-05

Table 3: Correlations with success and coordination.

Our first metric for identifying entrainment, Measure 1, the number of cues the speaker pair has in common, is negatively correlated with mean latency and proportion of interruptions, our two measures of poor coordination. Its correlation with score, though not significant, is positive. So, more entrainment in BPCs under Measure 1 means smaller latency before backchannels and fewer interruptions, while there is a tendency for such entrainment to be associated with higher scores.

Our second entrainment metric, Measure 2, captures the similarities between speaker means of the 10 features associated with BPCs. To test correlations of this measure with task success, we collapse the ten features into a single measure by taking the negated Euclidean distance between each speaker pair's 2 vectors of means; this measure tells us how close these speakers are across all features examined. Under this analysis, we find that Measure 2 is negatively correlated with mean latency and positively correlated with score. Both correlations are strong and highly significant. Again, the correlation with interruptions is negative, although not significant. Thus, more entrainment defined by this metric means shorter latency between turns, fewer interruptions, and again and more strongly, higher scores.

We thus find that, the more entrainment at the global level, the better the coordination between the partners and the better their performance on their joint task. These results provide evidence of the importance of BPC entrainment to dialogue.

5 Conclusion

In this paper we discuss the role of entrainment in turn-taking behavior and its impact on conversational coordination and task success in the Columbia Games Corpus. We examine a novel form of entrainment, entrainment in BPCs – characteristics of speech segments that are followed by backchannels from the interlocutor. We employ three measures of entrainment – two global and one local – and find evidence of entrainment in all three. We also find correlations between our two global entrainment measures and conversational coordination and task success. In future, we will extend this analysis to the complementary turn-taking category of turn-yielding cues and explore how a spoken dialogue system may take advantage of information about entrainment to improve dialogue coordination and the user experience.

6 Acknowledgments

This material is based on work supported in part by the National Science Foundation under Grant No. IIS-0803148 and by UBACYT No. 20020090300087.

References

L. Bell, J. Boye, J. Gustafson, and M. Wiren. 2000. Modality convergence in a multimodal dialogue system. In *Proceedings of 4th Workshop on the Semantics and Pragmatics of Dialogue (GOTALOG)*.

L. Bell, J. Gustafson, and M. Heldner. 2003. Prosodic adaptation in human-computer interaction. In *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*.

S.E. Brennan and H.H. Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(6):1482–1493.

S.E. Brennan. 1996. Lexical entrainment in spontaneous dialog. In *Proceedings of the International Symposium on Spoken Dialog (ISSD)*.

E.H. Buder, A.S. Warlaumont, D.K. Oller, and L.B. Chorna. 2010. Dynamic indicators of Mother-Infant Prosodic and Illocutionary Coordination. In *Proceedings of the 5th International Conference on Speech Prosody*.

R. Coulston, S. Oviatt, and C. Darves. 2002. Amplitude convergence in children's conversational speech with animated personas. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP)*.

J. Edlund, M. Heldner, and J. Hirschberg. 2009. Pause and gap length in face-to-face interaction. In *Proceedings of Interspeech*.

D. Goleman. 2006. *Social Intelligence: The New Science of Human Relationships*. Bantam.

A. Gravano and J. Hirschberg. 2009. Backchannel-inviting cues in task-oriented dialogue. In *Proceedings of SigDial*.

A. Gravano and J. Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, 25(33):601–634.

M. Heldner, J. Edlund, and J. Hirschberg. 2010. Pitch similarity in the vicinity of backchannels. In *Proceedings of Interspeech*.

A. Nenkova, A. Gravano, and J. Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proceedings of ACL/HLT*.

K. Niederhoffer and J. Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

M. J. Pickering and S. Garrod. 2004. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27:169–226.

D. Reitter, F. Keller, and J.D. Moore. 2006. Computational modelling of structural priming in dialogue. In *Proceedings of HLT/NAACL*.

E. Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences. In D. Tannen, editor, *Analyzing Discourse: Text and Talk*, pages 71–93. Georgetown University Press.

S. Stoyanchev and A. Stent. 2009. Lexical and syntactic priming and their impact in deployed spoken dialogue systems. In *Proceedings of NAACL*.

A. Ward and D. Litman. 2007. Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora. In *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education*.

N.G. Ward and S.K. Mamidipally. 2008. Factors Affecting Speaking-Rate Adaptation in Task-Oriented Dialogs. In *Proceedings of the 4th International Conference on Speech Prosody*.