Analyzing the Effect of Entrainment on Dialogue Acts

Masahiro Mizukami†, Koichiro Yoshino†, Graham Neubig†, David Traum‡, Satoshi Nakamura†
†Nara Institute of Science and Technology, Japan
‡USC Institute for Creative Technologies, USA
masahiro-mi@is.naist.jp

Abstract

Entrainment is a factor in dialogue that affects not only human-human but also human-machine interaction. While entrainment on the lexical level is well documented, less is known about how entrainment affects dialogue on a more abstract, structural level. In this paper, we investigate the effect of entrainment on dialogue acts and on lexical choice given dialogue acts, as well as how entrainment changes during a dialogue. We also define a novel measure of entrainment to measure these various types of entrainment. These results may serve as guidelines for dialogue systems that would like to entrain with users in a similar manner.

1 Introduction

Entrainment is a conversational phenomenon in which dialogue participants synchronize to each other with regards to various factors: lexical choice (Brennan and Clark, 1996), syntax (Reitter and Moore, 2007; Ward and Litman, 2007), style (Niederhoffer and Pennebaker, 2002; Danescu-Niculescu-Mizil et al., 2011), acoustic prosody (Natale, 1975; Coulston et al., 2002; Ward and Litman, 2007; Kawahara et al., 2015), pronunciation (Pardo, 2006) and turn taking (Campbell and Scherer, 2010; Beňuš et al., 2014). Previous works have reported that entrainment is correlated with dialogue success, naturalness and engagement.

However, there is much that is still unclear with regards to how entrainment affects the overall flow of the dialogue. For example, can entrainment also be observed in choice of dialog acts? Is entrainment on the lexical level more prevalent for utterances of particular dialogue acts? Does the level of entrainment increase as dialogue progresses?

If the answer to these questions is affirmative, it will be necessary to model entrainment not only on the lexical level, but also on the higher level of dialog flow. In addition, it will be necessary to adapt any entrainment features of dialogue systems to be sensitive to dialogue acts or dialogue progression. Modeling such entrainment phenomena appropriately has the potential to increase the naturalness of the conversation and open new avenues in human-machine interaction.

In this paper, we perform a study of entrainment in an attempt to answer these three questions. First, we observe the entrainment of dialogue acts, measuring whether the choice of dialogue acts synchronizes with that of the dialogue partner. For example, if one dialogue participant tends to ask questions frequently, we may hypothesize that the number of questions from the partner may also increase. Secondly, we examine lexical entrainment features given dialogue acts. It is known that dialogue acts strongly influence content of utterances, and we hypothesize that, in the same manner, dialogue acts may strongly influence the level of lexical entrainment. Finally, we examine the increase of entrainment as dialogue progresses. Previous work has discussed that entrainment can be observed throughout the whole dialogue, but it is unclear whether entrainment increases in latter parts of the dialogue. To measure this, we divide dialogues in half, and compare the entrainment of the former and latter halves.

Experimental results show that entrainment of dialogue acts does occur, indicating that it is necessary for models of dialogue to consider this fact. In addition, we find that the level of lexicon synchronization depends on dialogue acts. Finally, we confirm a tendency of entrainment increasing through the dialogue, indicating that dialogue systems may need to progressively adapt their models to the user as dialogue progresses.

Proceedings of the SIGDIAL 2016 Conference, pages 310–318,
Los Angeles, USA, 13-15 September 2016.

2 Related Works

2.1 Varieties of entrainment

As mentioned in the introduction, entrainment has been shown to occur at almost every level of human communication (Levitan, 2013), including both human-human and human-system conversation.

In human-human conversation, Kawahara et al. (2015) showed the synchrony of backchannels to the preceding utterances in attentive listening, and they investigated the relationship between morphological patterns of backchannels and the syntactic complexities of preceding utterances. Levitan et al. (2015) showed the entrainment of latency in turn taking.

In human-system conversation, Campbell and Scherer (2010) tried to predict user's turn taking behavior by considering entrainment. Fandrianto and Eskenazi (2012) modeled a dialogue strategy to increase the accuracy of speech recognition by using entrainment intentionally. Levitan (2013) unified these two works.

One of the most important questions about entrainment with respect to dialogue systems is its association with dialogue quality. Nenkova et al. (2008) proposed a score to evaluate the lexical entrainment in highly frequent words, and found that the score has high correlation with task success and engagement. This indicates that lexical entrainment has an important role in dialogue. In addition, it suggests that entrainment of lexical choice is probably affected by more detailed dialogue information, such as dialogue act.

2.2 Lexical Entrainment

The entrainment score which was proposed by Nenkova et al. (2008) is calculated by word counts in a corpus, and comparing between dialogue participants. Specifically, we calculate a uni-gram language model probability $\mathrm{P_{S_1}}(w)$ and $\mathrm{P_{S_2}}(w)$ based on the word frequencies of speakers S_1 and S_2, and calculate the entrainment score of word class V, $\mathrm{En}(V)$ as:

$$\mathrm{En}(V) = -\sum_{w \in V} \left| \mathrm{P_{S_1}}(w) - \mathrm{P_{S_2}}(w) \right|. \quad (1)$$

These entrainment scores have a range from -2 to 0, where higher means stronger entrainment. We calculate the average of these entrainment scores for the dialogue partner ($\mathrm{En_p}(V)$) and non-partners ($\mathrm{En_{np}}(V)$).

In detail, we can express this formula with word count $\mathrm{C}_{S_1}(w)$ and $\mathrm{C}_{S_2}(w)$, and all of words W as,

$$\mathrm{En}(V) = -\sum_{w \in V} \left| \frac{\mathrm{C}_{S_1}(w)}{\sum_{w_i \in W} \mathrm{C}_{S_1}(w_i)} - \frac{\mathrm{C}_{S_2}(w)}{\sum_{w_i \in W} \mathrm{C}_{S_2}(w_i)} \right|. \quad (2)$$

Nenkova et al. (2008) used following word classes as V.

25MFC: 25 Most frequent words in the corpus. The idea of using only frequent words is based on the fact that we would like to avoid the score being affected by the actual content of the utterance, and focus more on the way things are said. In addition, this filtering of highly frequent words removes any specific words (i.e. named entity, speaker's name) and words specific to the dialogue topic. This word class was highly and significantly correlated with task success in the previous work. We mainly used this word class in this paper.

25MFD: 25 Most frequent words in the dialogue. This word class was correlated with task success, like 25MFC.

ACW: Affirmative cue words (Gravano et al., 2012). This word class includes *alright, gotcha, huh, mm-hm, okay, right, uh-huh, yeah, yep, yes,* and *yup*. This class was correlated with turn-taking.

FP: Filled pauses. This word class includes *uh, um,* and *mm*. It was correlated with overlaps.

ACW and FP were pre-defined, but 25MFC and 25MFD are calculated from corpora considering frequency (V is a subset of W).

In order to use these measures to confirm whether entrainment is occurring between dialogue partners, these scores can be compared between the actual conversation partner, and an arbitrary other speaker from the database. If entrainment is actually occurring, then the score will be higher for the conversation partner than the score for the non-partner. Figure 1 shows an example of pairs used for calculation of these scores.

First, to confirm the results for previous work, we calculated the entrainment score of 25MFC using the Switchboard Corpus (Table 1). We can see that there is a difference of the entrainment score

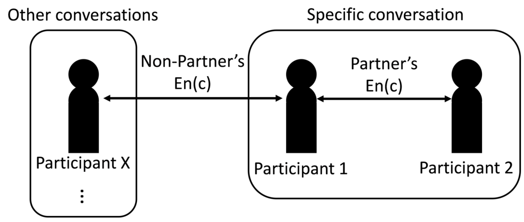

Figure 1: How to compare scores between the partner and non-partners

Table 1: The entrainment score of 25MFC

	Partner	Non-Partner
En(25MFC)	-0.211	-0.248

between "partner" who is talking the speaker and "non-partner" who is not talking with the speaker, as reported in previous work.

3 Extending the Entrainment Score

Our first contribution is an extension to the entrainment score that allows us to more accurately clarify the hypotheses that we stated in the introduction. This is necessary because the entrainment score given in Eqn. (2) does not consider the total size and variance of data to be calculated, and can be heavily influenced by data sparsity. This result in the score being biased when we compare target phenomena with different vocabulary sizes or data sizes.

For example, when considering the amount of entrainment that occurred for two different speakers, the entrainment score will tend to be higher for the more verbose speaker, regardless of the amount of entrainment that actually occurred. In addition, if we are comparing entrainment for two different sets of target phenomena, such as words and dialogue acts, the entrainment score will tend to be higher for the phenomenon that has a smaller vocabulary and thus less sparsity (in this case, dialogue acts). Thus, we propose a new "Entrainment Score Ratio" measurement that uses the rank in entrainment score, and language model smoothing to alleviate the effects of sparsity.

3.1 Entrainment Score Ratio

First, instead of using the entrainment score itself, we opt to use the relative position of the entrainment score of the partner compared to other non-partner speakers in the corpus. The entrainment score ratio is calculated according to the following procedure:

1. Calculate the entrainment score of the dialogue partner $\mathrm{En_p}(V)$. Also calculate entrainment scores of all non-partners in the corpus $\mathrm{En_{np_{1,\ldots,N}}}(V)$.

2. Compare the partner's entrainment score and all non-partners' entrainment scores.
$$\mathrm{Win}(\mathrm{En_p}(V), \mathrm{En_{np_i}}(V)) = \begin{cases} 1 & (\mathrm{En_p}(V) > \mathrm{En_{np_i}}(V)) \\ 0.5 & (\mathrm{En_p}(V) = \mathrm{En_{np_i}}(V)) \\ 0 & (\mathrm{En_p}(V) < \mathrm{En_{np_i}}(V)) \end{cases}$$

3. Calculate the ratio with which the partner's entrainment score exceeds that of the non-partners.
$$\mathrm{Ratio}(V) = \frac{1}{|N|} \sum_{i \in N} \mathrm{Win}(\mathrm{En_p}(V), \mathrm{En_{np_i}}(V))$$

Because this score is the ratio that dialogue with the partner takes a higher entrainment score than other combinations with non-partners, it is not sensitive to the actual value of the entrainment score, but only the relative value compared to non-partners. This makes it more feasible to compare between phenomena with different vocabulary sizes, such as lexical choice and dialogue act choice. While the entrainment score for dialogue acts may be systematically higher due to its smaller vocabulary size, the relative score compared to non-partners can be expected to be approximately equal if the effect of entrainment is the same between the two classes.

3.2 Dirichlet Smoothing of Language Models

While the previous ratio score has the potential to alleviate problems due to comparing different types of phenomena, it does not help with problems caused by comparing data sets with different numbers of data points. The reason for this is that the traditional entrainment score (Nenkova et al., 2008) used uni-gram probabilities, the accuracy of which is dependent on the amount of data used to calculate the probabilities. Thus for smaller data sets, these probabilities are not well trained, and show a lower similarity when compared with those of other speakers in the corpus. In order to create a method more robust to these size differences, we introduce a method that smooths these probabilities to reduce differences between distributions of different data sizes.

Specifically, the definition of a unigram distribution of a portion of the corpus (split by speaker s, dialogue act d, part of dialogue p) using maximum likelihood estimation is,

$$\mathrm{P}_{\mathrm{ML,s}}(w|d,p) = \frac{\mathrm{C}_s(w_{d,p})}{\sum_{w_{d,p} \in W_{d,p}} \mathrm{C}_s(w_{d,p})}. \quad (3)$$

When the size of data for speaker s is small, there will not be enough data to properly estimate this probability. To cope with this problem, we additively smooth the probabilities by introducing a smoothing factor α and large background language model $\mathrm{P}_{\mathrm{ML}}(w)$ which was trained using all of the available data:

$$\mathrm{P}_{\mathrm{DS,s}}(w|d,p) = \frac{\mathrm{C}_s(w_{d,p}) + \alpha \mathrm{P}_{\mathrm{ML}}(w)}{\sum_{w_{i,d,p} \in W_{d,p}} \mathrm{C}_s(w_{i,d,p}) + \alpha}. \quad (4)$$

This additive smoothing is equivalent to introducing a Dirichlet distribution conditioned on $\mathrm{P}_{\mathrm{ML}}(w)$ as a prior probability for the small language model distribution of $P_{DS,s}(w|d,p)$ (MacKay and Peto, 1995). We choose Dirichlet smoothing because it is a simple but effective smoothing method. We determine the hyperparameter α by defining a Dirichlet process (Teh et al., 2012) prior, and maximizing the likelihood using Newton's method[1].

To verify that this method is effective, we calculated averages and variances of the standard entrainment score and the entrainment score using this proposed smoothing technique (Table 2). From the results, we can see that the entrainment score rate for partners is slightly higher with smoothing, demonstrating that the smoothed scores are as effective, or slightly more effective in identifying the actual conversational partner. In addition, the difference between variances of entrainment scores has decreased, showing that smoothing has reduced the amount of fluctuation in scores. This indicates that the smoothing works effectively to reduce the negative influence of population size when we compare distributions that have different population sizes. Because of this, for the analysis in the rest of the paper we use this smoothed entrainment score.

[1]The scripts for this and other calculations will be public at the link below:
`https://github.com/masahiro-mi/entrainment`

4 Measured Entrainment Scores

In this section, we explain in detail the there varieties of entrainment that we examined.

4.1 Entrainment Score of Dialogue Acts

While entrainment of various phenomena has been reported in previous work, it is still not clear how entrainment affects the dialogue acts used by the conversation participants. The first thing we examine in this paper is the amount of entrainment occurring in dialogue acts, and the entrainment score of dialogue acts $\mathrm{En}(D)$ is calculated according to the differences in distributions of dialogue acts between dialogue participants. Frequency of each dialogue act $\mathrm{P}_{\mathrm{DS,S_1}}(d)$ and $\mathrm{P}_{\mathrm{DS,S_2}}(d)$ of each speaker S_1, S_2 for a certain dialogue act d is used in the following equation:

$$\mathrm{En}(D) = -\sum_{d \in D} \left| \mathrm{P}_{\mathrm{DS,S_1}}(d) - \mathrm{P}_{\mathrm{DS,S_2}}(d) \right|. \quad (5)$$

4.2 Lexical Entrainment Given Dialogue Acts

In the previous work, it is reported that there is an entrainment of lexical selection between dialogue participants. However, we can also hypothesize that such entrainment is more prominent for utterances with a particular dialogue act. For example, if one dialogue participant tends to say a specific backchannel frequently, the partner may change to use the same backchannel. On the other hand, when one dialogue participant has his/her own answer for a question, he/she will likely not borrow the words from the partner.

In order to examine this effect, we extended the entrainment score for lexical selection to evaluate an entrainment of lexical selection given the dialogue act of the utterance. The extended entrainment score $\mathrm{En}(c|d)$, the score for a lexical selection given a dialogue act, is defined by using conditional language model probabilities $\mathrm{P}_{\mathrm{DS,S_1}}(w|d)$ and $\mathrm{P}_{\mathrm{DS,S_2}}(w|d)$ of each speaker S_1 and S_2. Specifically, we define it as follows:

$$\mathrm{En}(V|d) = -\sum_{w \in V} \left| \mathrm{P}_{\mathrm{DS,S_1}}(w|d) - \mathrm{P}_{\mathrm{DS,S_2}}(w|d) \right|. \quad (6)$$

Using this measure, we clarify whether entrainment of lexicons has been affected by dialogue acts, and also which dialogue acts are more likely to be conducive to entrainment.

Table 2: The entrainment score variance with/without smoothing

		Partner		Non-Partner	
	Ratio(V)	Ave.	Var.	Ave.	Var.
w/o smoothing	0.671	-0.211	0.00537	-0.248	0.00181
w/ smoothing	0.706	-0.0983	0.00108	-0.123	0.000778

4.3 Increase of Entrainment through Dialogue

Nenkova et al. (2008) noted that the entrainment score between dialogue partners is higher than the entrainment score between non-partners in dialogue. While they reported the overall trend of the entrainment score throughout the dialogue, whether the level of entrainment changes throughout the dialog is also an important question, as it will indicate how dialogue systems must display entrainment properties to build a closer relationship with their dialogue partners. If entrainment is changing through a conversation, we can hypothesize that the entrainment score will be larger at the end of dialogue than the score at the start of dialogue.

We analyzed the extent of change in entrainment by splitting one dialogue into earlier and later parts. We calculated the entrainment score between dialogue participants in earlier/later parts of dialogue, and compared these scores.

5 Corpus

As our experimental data, we used the Switchboard Dialogue Act Corpus, which is annotated with dialogue acts according to the DAMSL standard (Discourse Annotation and Markup System of Labeling) (Jurafsky et al., 1997) for each utterance. The DAMSL has 42 types of dialogue act tags, while there were 220 tags used in the original Switchboard Corpus, Jurafsky et al. (1997) clustered the 220 tags into 42 rough-arained scale classes, and reported labeling accuracy of .80 according to the pairwise Kappa statistic.

This corpus consists of 302 male and 241 female speakers. The number of conversations is 1,155, and the number of utterances is 221,616. Each speaker is tagged with properties of sex, age, and education level.

Table 3: The entrainment score of dialogue acts

	Partner	Non-Partner	Ratio
DA	**-0.568****	-0.715	0.675

* $p < 0.10$, ** $p < 0.05$

6 Experimental Results

6.1 Entrainment of Dialogue Acts

First, we analyze the entrainment of dialogue acts based on the method of Section 4.1. We hypothessize that we can observe the entrainment of dialogue acts like other previously observed factors. To examine this hypothesis, we calculated the entrainment score of dialogue acts and compared between partner and non-partners. To measure the significance of these results, we calculated p-value of entrainment scores between partner and non-partner with the t-test.

Table 3 shows that there is a significant difference ($p < 0.05$) of entrainment score between partner and non-partner, with partners scoring significantly higher than non-partners. This result shows that the entrainment of dialogue acts can be observed in human-human conversation, and suggests that there may be a necessity to consider entrainment of dialogue act selection in human-machine interaction.

6.2 Lexical Entrainment given Dialogue Acts

Next, we analyze the entrainment of lexical choice given the 42 types of dialogue acts based on the method of Section 4.2. We can assume that the dialogue act affects the entrainment of lexicons, which indicates that entrainment scores are different depending on the type of the given dialogue act.

In addition, we calculate entrainment score rate and Cohen's d (Cohen, 1988) to evaluate the effect size. Cohen's d is standardized mean difference between two groups, and can calculate the amount that a particular factor effects a value while considering each group's variance. If these groups have a large difference, Cohen's d will be larger, with values less than 0.2 being considered small, values around 0.5 being medium, and values larger than 0.8 being considered large.

We show the result in Table 4, and emphasize scores that are over 0.5 in Cohen's d, and over 0.55 in Ratio(V).

We can first notice an increase of the entrainment score is more prominent given some dialogue acts. Entrainment is particularly prevalent for acts that have little actual informational content, such as greeting, backchannel, agree, answer, and repeating.

In addition, we focus on why Conventional Opening and Conventional Closing were increased in the entrainment score. This is because that Conventional Opening and Conventional Closing contain greetings ("hi", "hello") or farewells ("bye", "see you"), which show higher entrainment scores than other dialogue acts. It should be noted that this phenomenon of performing a fixed response to a particular utterance is also often called "coordination", and distinguished from entrainment. However, it is difficult to distinguish between entrainment and coordination definitely with our current measures, and devising measures to capture this distinction is future work.

On other hand, dialogue acts that express one's opinion such as Apology, Action-directive, Negative non-no answers, as well as some questions do not increase entrainment scores.

6.3 Change in Entrainment through Dialogue

In addition, we analyzed the increase of entrainment based on the method of Section 4.3. We calculated lexical entrainment scores of the earlier and later parts. "Earlier" is the entrainment score between utterances in the earlier part of dialogue, and "Later" is the entrainment score between utterances in the later part. We hypothesize that "Later" will have a higher entrainment score than "Earlier," as it is possible that dialogue participants will demonstrate more entrainment as they talk for longer and grow more comfortable with each other.

In addition, we calculate "Cross," the entrainment score between the earlier and the later parts of dialogue. We calculated this because we can also hypothesize that the effect of entrainment is delayed, and words spoken in the earlier part of the conversation may appear in the later part of the partner's utterances. Figure 2 shows the pairs used for the calculation. We show the result in Table 5.

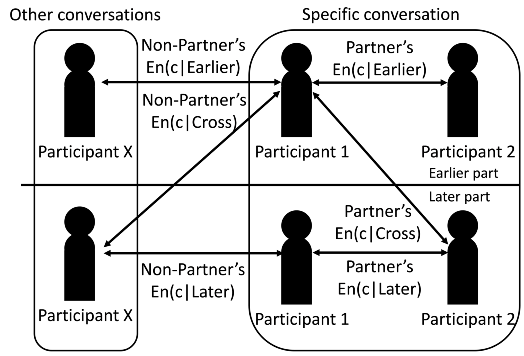

Figure 2: How we compare between earlier and later parts

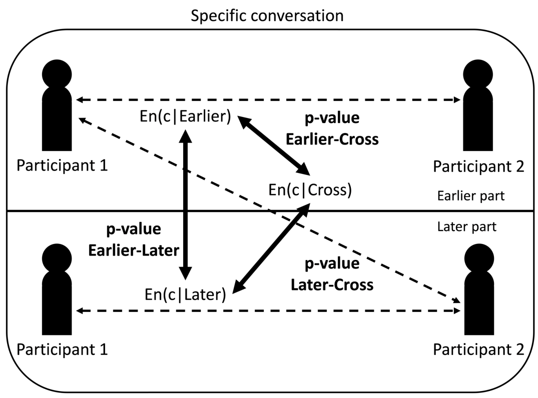

Figure 3: How to calculate p-values between each part in partner

From these results, we can see that there is a significant difference of entrainment score between partner and non-partner in all of the parts. This indicates that lexical entrainment can already be observed in the earlier part of dialogue.

In addition, we calculated p-values with the two-sided t test for partner entrainment scores between each part. Figure 2 shows an example of pairs used for calculation of p-values. We compare partner's entrainment scores between early, later, and cross, to indicate how the entrainment score changes in the partner through the dialogue. In fact, we compare three combinations of partner's entrainment scores, such as En(c|Earlier) and En(c|Later), En(c|Earlier) and En(c|Cross), and En(c|Later) and En(c|Cross). Table 6 shows that p-values of entrainment scores between each part in the partner. We find that the value of the entrainment score of the later part increased slightly over the entrainment score of the earlier part, but the increase was not significant. These results show that if there is a difference in entrainment between

Table 4: The entrainment score of lexicons given a dialogue act

	Partner $\mathrm{En_p}(V)$	Non-Partner $\mathrm{En_{np}}(V)$	Cohen's d	Ratio(V)
25MFC | Conventional-closing	**-0.0391****	-0.185	**1.50**	**0.703**
25MFC | Acknowledge (Backchannel)	**-0.201****	-0.252	**0.527**	**0.659**
25MFC | Statement-non-opinion	**-0.0930****	-0.113	0.434	**0.672**
25MFC | Statement-opinion	**-0.154****	-0.192	0.418	**0.634**
25MFC | Conventional-opening	**-0.0112****	-0.0370	0.406	0.542
25MFC | Segment (multi-utterance)	**-0.203****	-0.232	0.382	**0.618**
25MFC | Agree/Accept	**-0.279****	-0.325	0.367	**0.592**
25MFC | Appreciation	**-0.282****	-0.331	0.322	**0.564**
25MFC | Yes answers	**-0.320****	-0.375	0.274	**0.555**
25MFC | Non-verbal	**-0.104****	-0.124	0.259	**0.557**
25MFC | Abandoned or Turn-Exit, Uninterpretable	**-0.203****	-0.228	0.244	**0.592**
25MFC | Hedge	**-0.170****	-0.191	0.132	0.532
25MFC | Wh-Question	**-0.147****	-0.160	0.122	0.530
25MFC | Backchannel in question form	**-0.134****	-0.152	0.118	0.528
25MFC | No answers	**-0.199****	-0.220	0.118	0.523
25MFC | Rhetorical-Questions	**-0.0644****	-0.0754	0.102	0.522
25MFC | Response Acknowledgement	**-0.207****	-0.227	0.100	0.521
25MFC | Repeat-phrase	**-0.115****	-0.128	0.0962	0.522
25MFC | Other	-0.160	**-0.150****	0.0772	0.476
25MFC | Quotation	**-0.0817****	-0.0905	0.0749	0.517
25MFC | Collaborative Completion	**-0.0867****	-0.0929	0.0616	0.514
25MFC | Yes-No-Question	**-0.223***	-0.227	0.0490	0.512
25MFC | Hold before answer/agreement	**-0.104****	-0.112	0.0488	0.511
25MFC | Summarize/reformulate	**-0.109****	-0.114	0.0380	0.512
25MFC | Signal-non-understanding	**-0.0377****	-0.0404	0.0377	0.507
25MFC | Declarative Yes-No-Question	-0.134*	-0.138	0.0348	0.512
25MFC | Other answers	-0.0584*	-0.0620	0.0313	0.507
25MFC | Maybe/Accept-part	-0.0204	-0.0221	0.0247	0.503
25MFC | Self-talk	-0.0189	-0.0205	0.0235	0.503
25MFC | Thanking	-0.0180	-0.0195	0.0227	0.502
25MFC | Reject	-0.0670	-0.0696	0.0209	0.504
25MFC | Negative non-no answers	-0.0600	-0.0581	0.0181	0.497
25MFC | Open-Question	-0.0877	-0.0894	0.0166	0.504
25MFC | Affirmative non-yes answers	-0.134	-0.136	0.0161	0.504
25MFC | Downplayer	-0.0238	-0.0247	0.0111	0.501
25MFC | Declarative Wh-Question	-0.0147	-0.0152	0.00797	0.501
25MFC | Action-directive	-0.0935	-0.0944	0.00748	0.502
25MFC | Dispreferred answers	-0.0514	-0.0522	0.00716	0.502
25MFC | Apology	-0.0183	-0.0179	0.00667	0.500
25MFC | 3rd-party-talk	-0.00969	-0.00955	0.00369	0.500
25MFC | Offers, Options Commits	-0.0204	-0.0205	0.00222	0.500
25MFC | Or-Clause	-0.0502	-0.0502	0.000816	0.500

N(Number of target speaker) = 2310, * $p < 0.10$, ** $p < 0.05$

Table 5: The entrainment score for combinations of part

	Partner	Non-Partner	Rate
En(25MFC|Earlier)	**-0.106****	-0.126	0.658
En(25MFC|Cross)	**-0.106****	-0.127	0.666
En(25MFC|Later)	**-0.104****	-0.126	0.674

* $p < 0.10$, ** $p < 0.05$

Table 6: The p-values for partner's entrainment score between each part

		p-value
En(25MFC|Earlier)	En(25MFC|Later)	0.222
En(25MFC|Earlier)	En(25MFC|Cross)	0.238
En(25MFC|Later)	En(25MFC|Cross)	0.00425

earlier and later parts of the conversation, the difference is slight.

7 Conclusion

In this paper, we focused on the entrainment with respect to dialogue acts and dialogue progression, and analyzed for three phenomena: the entrainment of dialogue acts, the entrainment of lexical choice given dialogue acts, and the change in entrainment as dialogue progresses.

From the results, we found that the entrainment of dialogue acts was observed in conversation. Within dialogue systems, this has the potential to contribute to modelling of dialogue strategy, and potentially allow the system to have a closer relationship with the partner.

We also found that lexical entrainment has a different tendency depending on the dialogue act of the utterance. This has the potential to contribute to models of language generation, which can consider entrainment of each dialogue act.

Finally, we analyzed the differences of entrainment depending on the part of the dialogue. From results, we found that there is either only a slight effect, or no effect of the part of the dialogue under consideration.

In future works, we will try an analysis of the entrainment in dialogue that considers the effect of coordination.

Acknowledgement

This work is supported by JST CREST.

References

Štefan Beňuš, Agustín Gravano, Rivka Levitan, Sarah Ita Levitan, Laura Willson, and Julia Hirschberg. 2014. Entrainment, dominance and alliance in supreme court hearings. *Knowledge-Based Systems*, 71:3–14.

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

Nick Campbell and Stefan Scherer. 2010. Comparing measures of synchrony and alignment in dialogue speech timing with respect to turn-taking activity. In *INTERSPEECH*, pages 2546–2549.

Jacob Cohen. 1988. Statistical power analysis for the behavioral sciences. 2nd edn. hillsdale, new jersey: L.

Rachel Coulston, Sharon Oviatt, and Courtney Darves. 2002. Amplitude convergence in children's conversational speech with animated personas. In *Proc. ICSLP*, volume 4, pages 2689–2692.

Cristian Danescu-Niculescu-Mizil, Michael Gamon, and Susan Dumais. 2011. Mark my words!: linguistic style accommodation in social media. In *Proceedings of the 20th international conference on World wide web*, pages 745–754. ACM.

Andrew Fandrianto and Maxine Eskenazi. 2012. Prosodic entrainment in an information-driven dialog system. In *INTERSPEECH*, pages 342–345.

Agustín Gravano, Julia Hirschberg, and Štefan Beňuš. 2012. Affirmative cue words in task-oriented dialogue. *COLING*, 38(1):1–39.

Dan Jurafsky, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*, pages 97–102.

Tatsuya Kawahara, Takashi Yamaguchi, Miki Uesato, Koichiro Yoshino, and Katsuya Takanashi. 2015. Synchrony in prosodic and linguistic features between backchannels and preceding utterances in attentive listening. In *APSIPA*, pages 392–395. IEEE.

Rivka Levitan, Stefan Benus, Agustın Gravano, and Julia Hirschberg. 2015. Entrainment and turn-taking in human-human dialogue. In *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*.

Rivka Levitan. 2013. Entrainment in spoken dialogue systems: Adopting, predicting and influencing user behavior. In *HLT-NAACL*, pages 84–90.

David JC MacKay and Linda C Bauman Peto. 1995. A hierarchical dirichlet language model. *Natural language engineering*, 1(03):289–308.

Michael Natale. 1975. Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Journal of Personality and Social Psychology*, 32(5):790.

Ani Nenkova, Agustín Gravano, and Julia Hirschberg. 2008. High frequency word entrainment in spoken dialogue. In *Proc. ACL*, pages 169–172. Association for Computational Linguistics.

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

Jennifer S Pardo. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

David Reitter and Johanna D Moore. 2007. Predicting success in dialogue. In *Proc. ACL*.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. 2012. Hierarchical dirichlet processes. *Journal of the american statistical association*.

Arthur Ward and Diane Litman. 2007. Measuring convergence and priming in tutorial dialog. In *University of Pittsburgh*.