Beyond Static Measures: Temporal Analysis of Lexical Alignment in Human-Human Learning With a Teachable Robot

Anonymous Author

Anonymous Institute

Abstract. Lexical alignment occurs when conversational partners converge on similar linguistic patterns, fostering engagement and rapport. Traditional approaches to alignment computation often focus more on the summary statistics computed at the end of the conversation, which usually do not capture the conversational dynamics efficiently. In this work, we investigate how alignment evolves in a conversation by modeling lexical alignment between human dyads with a logarithmic curve in human-human-robot teaching interactions. We then evaluate how the properties of the fitted curves relate to partner rapport and learning gain. We find that along with the summary statistics, the curve parameters and the time taken to reach key alignment moments significantly predict partner rapport, though their impact on learning gains requires further investigation. We further found significant relationships between the early turns in the conversation and the overall alignment trajectories, indicating the importance of modeling conversational dynamics to plan real-time interventions in the robot, altering the alignment trajectories, and, consequently interaction outcomes.

Keywords: Lexical Alignment · Alignment Trajectories · Logarithmic Modeling.

1 Introduction

Lexical alignment is a phenomenon where interlocutors adopt similar vocabulary for the same underlying concepts [21], leading to a higher overlap in their shared mental models [15]. This phenomenon is extremely prevalent in everyday conversation, where humans tend to align their lexical choices, speech properties, and even non-verbal cues such as gestures [1] to facilitate mutual understanding. This phrase-level alignment has been shown to influence the increase in engagement [14] and task success rates in human-human task-oriented dialogues [8,9,16,18,25]. This effect is not restricted to Human-Human dialogues and is prevalent in human-agent [12,22] and human-robot settings [2], affecting agent comprehension and rapport. In many dialogue studies, alignment is typically reduced to a single, summary measure, for example, the final turn's alignment value to represent convergence [2,5], thus missing the rich dynamic patterns unfolding over the course of a conversation.

In our work, we extend these ideas into the realm of human–human-robot interactions in a collaborative learning context investigating the continuous evolution of alignment or the *entrainment curve* in these conversations. Our study is set in a collaborative learning scenario where two students work together with a teachable NAO robot to solve ratio problems. We posit that not only does the final alignment score matter but that the entire trajectory of alignment throughout a dialogue may be more informative regarding certain learner behaviors such as partner rapport and learning gain. Moreover, by modeling these continuous patterns, we gain the potential to identify early indicators of learners' behaviors, thereby enabling real-time interventions. For example, in our context, the robot can continuously model these human behaviors at the beginning of the conversations and then intervene to influence the final outcomes.

To explore these ideas, we first assess whether classical metrics of alignment such as the final turn's value, total turns, and mean alignment scores are effective predictors of learning gain and partner rapport. Next, we fit a logarithmic model to the entrainment values throughout the conversation, allowing us to extract properties of the curve, e.g., rate of convergence, intercept, slopes, and time taken to reach certain moments in entrainment and evaluate their relationship with the variables of interest. Finally, we fit curves using a small subset of the whole entrainment data and assess the predictive power of these early curves against the complete trajectory, asking whether early behavior can forecast the ultimate alignment pattern and related performance outcomes. Although our learning setting is a multiparty interaction, we focus our analysis on the human–human part of this multiparty interaction for this paper, which we hypothesize is significant for the final conversational outcomes. We aim to answer the following research questions:

1. **RQ1**: Do the properties of entrainment curves add more predictive information about learner behaviors than the final alignment score alone? If so, which properties of the entrainment curve most strongly reflect learner engagement and rapport throughout the conversation?
2. **RQ2**: How can early conversation data be used to predict the complete entrainment trajectory and, in turn, forecast overall learning gain and partner rapport?

1.1 Background

Shorten this section by removing lexical alignment definitions. Just keep the methods

Lexical Alignment Methods Early methods for assessing alignment focused on lexical priming, measuring how often a word introduced by one speaker is repeated by another within a specific time window [26]. These approaches often relied on linear regression to model how repetition patterns changed over time. However, they primarily captured short-term alignment effects and failed to account for broader temporal trends. Another widely used metric, vocabulary alignment, quantified convergence by analyzing the proportion of shared word usage between speakers [11]. Higher lexical convergence, particularly in high-frequency words, has been linked to more natural conversational flow and improved task success. More advanced statistical frameworks like the Hierarchical Alignment Model [4] exhibit modeling capabilities for linguistic structure repetitions, offering a more nuanced perspective on conversational adaptation. Some recent works have explored neural models to capture conceptual alignment, moving beyond direct word repetition to identify overarching themes and contextual structures in dialogue [21] [27] [24] [20].

Lexical alignment has also been examined in dyadic and multiparty interactions, particularly in collaborative education dialogues and human-robot interactions [21] [2] [6] [17]. While studies have demonstrated that alignment enhances communication and rapport in these settings, the complexity of task-based human-human-robot dialogues where relationships among dyads vary dynamically and have left significant gaps in our understanding of alignment mechanisms in these interactions [2].

A crucial limitation of existing alignment models is their retrospective nature in analyzing alignment after a conversation has occurred. However, understanding how alignment evolves over time could offer valuable predictive insights. In this work, we explore entrainment curves, a continuous representation of alignment patterns throughout a conversation, and investigate their ability to predict key conversation outcomes, such as rapport and learning gains. Furthermore, if these curves can be reliably estimated early in a conversation, they could enable real-time interventions, allowing for proactive modifications to conversational strategies. This, in turn, could influence final alignment patterns and improve communication effectiveness.

2 Methodology

2.1 System and Data

We analyze human-human conversations collected from interactions between pairs of human participants and a teachable NAO[1] robot. All dialogues and speaker information were collected from two different studies, both designed to examine how participants engage in collaborative problem-solving while instructing a teachable robot. The robot played an active role in the interaction, asking follow-up questions to clarify the students' explanations and build their understanding.

The first study involved 40 participants (35 males, 5 females, mean age: 19.64(1.25)) recruited from an undergraduate program at a public university in the USA. The students identified as 17 White, 13 Asian, 5 Black, 1 Latino, and 4 no answer. Only 28 of them worked in dyads and the rest of them worked alone with the robot. We only consider the data from the 14 dyads in this work. The experiment was conducted remotely via Zoom, with the robot and each participant appearing in separate windows. Participants collaborated with their

[1] https://www.robotlab.com/store/nao-power-v6-educator-pack

partner and the robot to solve ratio word problems, explaining each step to the robot over a 30-minute session. Of the 14 dyads one was excluded because of a complete lack of responses from one of the human speakers.

The second study was conducted in a controlled lab setting, where participants engaged in face-to-face interactions. The students worked in Human-Human-Robot groups and collaborated to solve math problems. Similar to the first study, participants instructed the robot while working together on the task. This study consisted of 28 participants (9 males, 18 females, 1 non-binary, mean age: 20.67(2.38)) recruited from an undergraduate program at a public university in the USA. The students identified as 13 Asian, 2 Black, 2 Nigerian, 10 White, and 1 Multiracial. We only consider the data from 13 dyads as the other 2 students worked with a facilitator in the study due to the absence of their partner.

All interactions were video recorded and system logs were captured to track participants' progress on the task. The collected data was manually transcribed and each utterance was annotated with speaker and receiver labels by one of the authors. Both studies included pre-tests and post-tests. These tests contained questions to evaluate the learning as well as additional conversation metrics, such as rapport with both the partner and the robot.

Both studies followed ethical guidelines approved by the Institutional Review Board of our Institution.

2.2 Alignment Computation and Variables of Interest

We computed lexical alignment using the concept of *shared expressions*, which are recurring text patterns identified at the utterance level and produced by both speakers in a dialogue [5]. An expression in a dyadic conversation (with speakers S1 and S2) becomes a "shared expression" when it is introduced by speaker S1 and then reused by speaker S2 later in the conversation. We used this notion of repeated expressions to devise our metric for alignment computation.

To quantify each speaker's contribution to the alignment process, we calculated the speaker's average repetition rate. For a given speaker S, the individual repetition rate is defined as:

$$RepetitionRate_{S_t} = \frac{\text{Number of Repeated Expressions by S till turn t}}{\text{Total number of n-grams by S till turn t}} \tag{1}$$

where n-grams represent contiguous sequences of words in the dialogue, up to n consecutive words. This measure captures how frequently a speaker *reuses expressions*, providing insights into their lexical alignment behavior. This measure would change as the conversation proceeds or conversation turns increase, as students will add more phrases to the overall shared vocabulary. Before calculating the n-grams we also preprocessed all the conversation data to make it lowercase and removed all the punctuations. Further, we limited the n-grams to a maximum sequence of 3 words in a phrase. So, for example, if speaker A has an utterance "*Hi. How are you?*", the n-grams generated would be '*hi*', '*how*', '*are*', '*you*', '*hi how*', '*how are*', '*are you*', '*hi how are*', and '*how are you*'. After generating these n-grams we removed all the *stop words* based on the popular lists of *stop words*[2] for the English language.

In our study, we computed this measure separately for Speaker A and Speaker B for every turn in the conversation, allowing us to analyze each speaker's role in lexical alignment. To compute an overall alignment at any specific turn(instance) in the dialogue, we computed a weighted average of both speakers' contributions:

$$\text{AlignmentScore}_t = \frac{RepetitionRate_{A_t} \times NGrams_{A_t} + RepetitionRate_{B_t} \times NGrams_{B_t}}{NGrams_{A_t} + NGrams_{B_t}} \tag{2}$$

where $RepetitionRate_{A_t}$ and $RepetitionRate_{B_t}$ represent the respective contributions of Speaker A and Speaker B till turn t and $NGrams_{A_t}$ and $NGrams_{B_t}$ are the number of n-grams till turn t by speaker A and speaker B respectively. This metric serves as a proxy for lexical alignment over time, capturing how alignment evolves across the dialogue. Higher values indicate a greater number of *shared expressions*, which can correlate with stronger lexical entrainment.

To perform temporal alignment analysis, the main contribution of our work, we developed a custom analysis system inspired by [7], incorporating several enhancements, such as stop word removal to filter out low-information words, n-gram matching with flexible window sizes, adjustable context windows, and word-specific alignment.

Rapport and Learning Gain. For this paper, we were mostly interested in evaluating how lexical entrainment between humans in a human-human-robot setting affects social and cognitive outcomes. We measure these using perceived partner rapport and learning gain. Each student in the dyad filled a post-survey on a six-point Likert scale, ranging from strongly disagree to strongly agree, with questions about perceived partner rapport. These questions were based on the rapport measures used in the previous literature [13] [10] [19] [23]. We averaged the value of all these answers to obtain a rapport metric for a student, which was then averaged again to generate the *average partner rapport* metric for the dyad. The mean and SD of *average partner rapport* were 5.10(0.49) in Study 1 and 5.11(0.40) in Study 2. An independent samples t-test between these values indicated no statistically significant difference.

For learning, each student has a pre-test and a post-test with a set of ratio problems. We adopted a normalized learning gain as a learning metric based on a student's pre-test and post-test scores. The following formula was used for this learning gain computation:

$$Gain = \begin{cases} \frac{\text{post - pre}}{\text{100 - pre}}, \text{post} > \text{pre} \\ 0, \text{post} = \text{pre} \\ \frac{\text{post - pre}}{\text{pre}}, \text{post} < \text{pre} \end{cases} \tag{3}$$

where *post* and *pre* are the post-test and pre-test scores respectively. We average these scores for both the students in the dyads to generate our *average normalized*

[2] https://gist.github.com/sebleier/554280

learning gain metric. The mean and SD of *average normalized learning gain* were 0.39(0.29) in Study 1 and 0.28(0.29) in Study 2. An independent samples t-test between these values indicated no statistically significant difference.

2.3 Analysis Framework

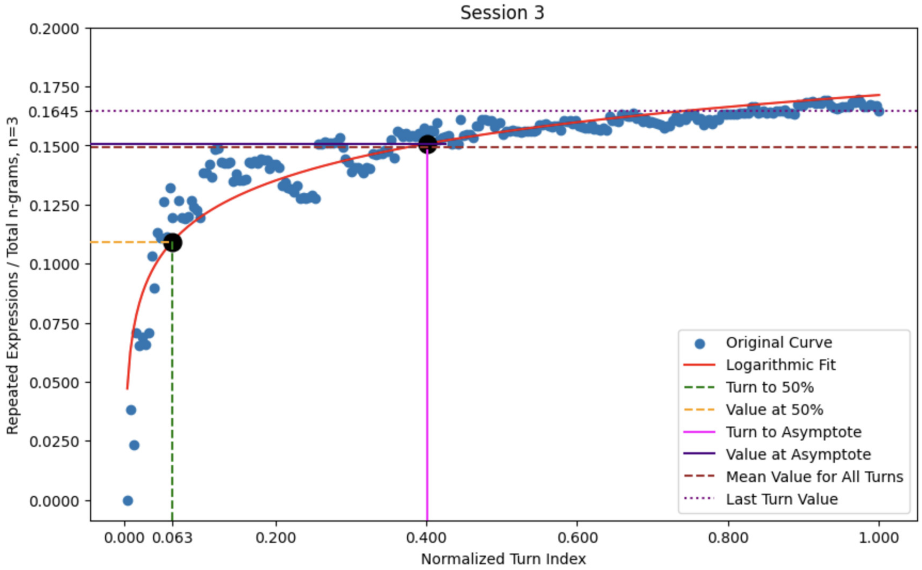
We perform our analysis to study lexical entrainment in two setups: Classical Analysis and Temporal Analysis. Classical analysis captures lexical alignment using a single summary metric at the end of the conversation cite papers, whereas Temporal Analysis examines entrainment dynamics and how it evolves by fitting an entrainment curve to the conversationcite papers. We also compare both methods in terms of their predictive power to capture the partner rapport and learning gain. This comparison can provide insights into choosing one over the other methods across different contexts.

Classical Analysis The classical analysis quantifies lexical entrainment using a single alignment score computed at the end of the conversation. This method has been widely used in prior literature to measure overall alignment tendencies in dialogues, treating it as a summary metric that reflects the final state of adaptation between speakers. For this paper, we adopt the following metrics to compute the final alignment score:

1. **Last Turn Value**: This metric uses the lexical alignment score computed at the last turn of the conversation while taking the entire conversation as a context. We use the cumulative value of this metric for both the speakers as an alignment score for the conversation. Several works have used this alignment score to describe overall lexical adaptation, particularly in studies focusing on collaborative learning with robots [2]. Additionally, studies have found that these final alignment measures can predict learning and collaboration success [18] [9], reinforcing their utility in conversational analysis.
2. **Mean Value**: This metric computes the mean of the lexical alignment score computed at every turn in the conversation, with only the turns preceding a particular turn in the conversation as the context. This provides a broader perspective on lexical entrainment by averaging alignment throughout the dialogue rather than focusing solely on the endpoint. Again, we use the cumulative value of this metric for both the speakers as an alignment score for the conversation. This metric can inform us how much an individual turn contributes to the overall alignment.
3. **Total Turns**: This metric uses the total number of turns in a conversation as a proxy for the alignment score. We use this metric to evaluate whether the length of the conversations is important in determining the final outcome. We hypothesize as the number of turns increases the repeated expressions will increase making the alignment score go up.

Despite their effectiveness in capturing alignment trends, these summary metrics fail to account for the nuances of entrainment dynamics throughout the

Fig. 1. The figure shows the alignment and the fitted logarithmic curves for one dyad with points marked to reflect: Mean and Last Turn Values(Actual curve) and Turn to 50%, Value at 50%, Turn to Asymptote, and Value at Asymptote (Fitted curve).



interaction. These scores overlook how lexical entrainment develops, whether there are key moments of entrainment shifts, and whether certain fluctuations in alignment are more representative of the underlying interaction behaviors. Figure 1 (see Original Curve) shows an example of an entrainment curve extracted from one of the sessions of our Human-Human data. The final alignment does not capture intermediate inflections in the curve, missing critical points where alignment may have increased or decreased in response to specific conversational events. Similarly, the mean score normalizes the behavior toward the center, smoothing out early inflection that again could represent certain key moments in the interaction, e.g., problem shift in our context.

Temporal Analysis We examine lexical entrainment as a dynamic process by fitting an entrainment curve to the conversation and analyzing its properties.

Linear Logarithmic Entrainment. For human-human conversations, we observed that lexical entrainment typically follows a *logarithmic pattern* (see figure 1), where alignment increases rapidly in the early turns and then stabilizes over time. This pattern suggests an initial period of high adaptation, after which speakers reach relatively stable entrainment levels.

To model this dynamic, we fit a *linear logarithmic curve* (Equation 4) to the alignment score over time.

$$E(t) = a.ln(t) + b \tag{4}$$

where:

- $E(t)$ represents the entrainment score at turn t,
- a captures the rate of lexical adaptation,
- b is the baseline alignment level, and
- $ln(t)$ implies the decreasing rate of change in the entrainment over time.

To ensure comparability across different sessions (mostly of different lengths), we normalize the number of turns to bring all the conversations on the same scale. This was done to rule out the effect of too short or too long conversations. This logarithmic function effectively models the initial sharp rise and subsequent stabilization in lexical entrainment. Any deviations from this fit in the actual conversation could indicate moments reflecting disruptions or potential shifts in the topic, e.g. problem context switch in our settings. We looked at the curve parameters, which tell about the growth rate and the stabilization level of alignment, and evaluated how well they can predict the outcomes of rapport and learning.

Apart from the curve parameters, we also look at the following properties of the curve:

1. Pseudo Inflection Point (Midpoint behavior): A logarithmic curve does not have a true inflection point (since its second derivative never changes sign), so we approximate an important transition zone where the growth starts to slow down significantly. We call this point a pseudo-inflection point and approximate it as half of its asymptotic value (maximum value in bounds). We then evaluate the behavior of these curves to reach this inflection point, the time to reach this point (or *Turn to 50%*), and the entrainment value at this point (or *Value at 50%*). This point is important as this marks a significant transition when the alignment shifts from rapid growth to a more stable, slow-changing state. This point can be further used to compare different conversations and groups, e.g. if alignment happens early or if it takes a long time. Conversations, where this midpoint occurs early, could suggest fast adaptation, while a later midpoint could indicate slower alignment, maybe due to different conversational strategies. Based on equation 5, we calculate these values as follows:

$$E(t_{50}) = \frac{E_{max} + E_{min}}{2} \tag{5}$$

$$t_{50} = e^{\frac{E(t_{50})-b}{a}} \tag{6}$$

   where E_{min} is the starting value of the fitted curve and E_{max} is the value of the curve at the last normalized turn. We evaluate whether these metrics, *Turn to 50%* and *Value at 50%*, can either individually or jointly determine the partner rapport and learning gain.
2. Asymptote Point (Saturation Behavior): A logarithmic curve theoretically grows indefinitely but at an ever-decreasing rate. However, for our conversational scenario, the curve tends to form a plateau near higher turn numbers.

This is an important zone for alignment as it represents a stable state (saturation). For our analysis, we define the asymptote as the point when the slope is 1% of the starting slope of the curve, meaning that the growth is extremely slow at this point and the curve almost reaches an asymptote. We then evaluate the behavior of these curves to reach this asymptote point, the time to reach this point (or *Turn to Asymptote*), and the entrainment value at this point (or Value at Asymptote). This point is important as it can tell if a certain conversation had the potential to reach even higher entrainment values if the conversation had lasted a bit longer. We see this in our results, where for some of the sessions this asymptote point lies outside the conversation duration, indicating that the alignment could have increased if the conversation lasted longer. Based on equation 7, we calculate these values as follows:

$$slope_{asymptote} = 0.01 \times \frac{a}{(1/\text{total turns})} \tag{7}$$

$$t_{asymptote} = \frac{a}{slope_{asymptote}} \tag{8}$$

$$E(t_{asymptote}) = a.ln(t_{asymptote}) + b \tag{9}$$

where a and b are the parameters of the fitted curve, and *total turns* represent the total number of utterances in the session. We evaluate whether these metrics, *Turn to Asymptote* and *Value at Asymptote*, can either individually or jointly determine the partner rapport and learning gain.

3. Area Under the Curve (AUC): AUC refers to the area under the fitted logarithmic curve and captures the cumulative effect of each turn in the conversation over time. We calculate this AUC by integrating the equation of our fitted model and generating the final value of AUC as shown in equation 11:

$$AUC_t = (a \times t \times ln(t)) - (a \times t) + (b \times t) \tag{10}$$

$$AUC = AUC_1 - AUC_{1/\text{total turns}} \tag{11}$$

where AUC_t is the AUC at the turn t, a and b are the curve parameters, *total turns* represent the total number of utterances in the session, and $ln(t)$ is the natural logarithm for turn t. AUC is useful for understanding the overall contribution of lexical alignment and is less sensitive to short-term fluctuations. We assess if AUC can better represent rapport and learning.

4. Reaching from Inflection to Asymptote: We also compare if the behaviors from the inflection point to asymptote differ across sessions and can explain the variance in rapport and learning. For this we used 2 metrics, (i.) the difference between *Turn to Asymptote* and *Turn to 50%*, and (ii.) the difference between *Value at Asymptote* and *Value at 50%*. These values correspond to late adaptation behaviors in learners and can indicate if they impact the final outcomes.

5. Comparison between Early Curves and Complete Curves: To take advantage of the dynamic nature of the conversation and see if the early trends in the conversation (e.g., first few turns) can predict the behaviors across

the whole conversation, we fit a new set of logarithmic curves, called early curves. We experimented with different numbers of turns but then used 25 turns as a dataset to build these curves. This 25 usually represents the end of the first problem boundary across many sessions. However, we did have sessions with a slightly longer or shorter number of turns in first problem boundaries. But to standardize the amount of turns across sessions, we only took the first 25 turns of each session. The mean turns in the sessions were 140.65(SD = 73.28), so 25 turns usually accounted for around 18% of the conversation. We then calculated direct correlations between the parameters of these early curves and the original curves. We aimed to find trends in the early conversation that can determine the outcome. Building these curves early in the conversation provides opportunities for interventions, that can alter the entrainment trajectories, hence altering the final outcomes.

3 Results

We evaluated if the study type had any effect on rapport and learning by running paired t-tests across the studies. None of the tests suggested significant differences between the distributions of these outcomes, hence we combined the samples from both the studies, resulting in a conversation dataset of 26 dyads. We also calculated Pearson's correlation between combined Average Partner Rapport and Average Normalized Learning Gain. No statistically significant correlation was found between these.

Our analyses have a combination of correlations and regressions based on the number of predictor variables. Every fitted regression will have the following form:

$$IV = a0 + a1 * P1 + a2 * P2 + a3 * P3 + \dots \tag{12}$$

where $P1$, $P2$, $P3$, $\dots$ are predictors, $a1$, $a2$, $a3$, $\dots$ are respectively the coefficients of these predictors, and $a0$ is the constant term, and IV is the dependent variable.

3.1 Classical Analysis

We use the *Last Turn* value, *Mean* value, and the *Total Turns* as a proxy for the alignment score in this analysis. Table 1 shows the descriptives for these metrics across all the dyads. We then calculate pairwise Pearson's correlation between all of these metrics and the Average Partner Rapport and Average Normalized Learning Gain. Table 1 shows the results(r and p-value) of these correlations. Last Turn (p-value < 0.05), Mean (p-value < 0.05) and Total Turns (p-value $<$ 0.01) have a strong statistically significant correlation to Average Partner Rapport. However, there is no relationship between these predictors and the Average Normalized Learning Gain. We further evaluated the relationships between the pair of these predictors using Pearson's correlations. All of these metrics are strongly correlated with each other with extreme statistical significance (p-value < 0.01) (see table 1).

Table 1. The table describes the analysis with the Classical Metrics: Total Turns, Last Turn Value, and Mean Value, and their correlations with each other and the final outcomes of rapport and learning.

Classical Analysis Descriptives			
	Total Turns	*Last Turn Value*	*Mean Value*
Mean(SD)	140.65(73.28)	0.17(0.05)	0.13(0.04)
Pearson's Correlation b/w Classical Metrics and Final Outcomes			
	Total Turns *r(p-value)*	*Last Turn Value* *r(p-value)*	*Mean Value* *r(p-value)*
Partner Rapport	0.52**(0.007)	0.42*(0.03)	0.43*(0.03)
Learning Gain	0.02(0.92)	0.04(0.85)	-0.06(0.79)
Pearson's Correlation b/w pairs of Classical Metrics			
Mean Value & Last Turn Value	0.96**(0.00)		
Last Turn Value & Total Turns	0.65**(0.0003)		
Mean Value & Total Turns	0.69**(0.00)		

3.2 Temporal Analysis

For temporal analysis, we fit a separate linear logarithmic curve for each session representing the entrainment behavior for each dyad. Table 2 shows the descriptive statistics of the Root-mean squared error (RMSE), the R^2 values, the slope values (a), and the intercept values(b) of the fitted curves across the dyads. An R^2 value close to 1 and an RMSE value close to 0 represents a good fit. We then calculate the relationship between these fitted curve parameters and the final outcome variables, Average Partner Rapport and Average Normalized Learning Gain using Pearson's correlation. Only RMSE has a strong statistically significant correlation with the Average Partner Rapport. There is no relationship between these parameters and the Average normalized Learning Gain.

We then evaluated if the curve parameters (a and b) have a joint relationship with our final outcomes. We fit two regression models using the equation 12 with a and b as $P1$ and $P2$ respectively, and Partner Rapport and Learning Gain as IV. Table 2 shows the results of these regressions. a and b have a joint significant relationship with Partner Rapport, with the whole model and all the regression coefficients being statistically significant (p-value of regression < 0.05, p-value of coefficient of $a < 0.05$, p-value of coefficient of $b < 0.01$). There is no joint relationship between these parameters and Learning Gain. We do see high coefficients but the model and the coefficients are not statistically significant.

We further assessed the relationships between the curve properties (as described in section 2.3) with Learning Gain and Partner Rapport using Pearson's correlation. Table 2 displays the descriptives of these curve properties followed the results of this correlation analysis. *AUC*, *Turn to 50%*, and *Value at 50%* show a strong statistically significant correlation with Partner Rapport. There is no other statistically significant relationship between curve properties and Partner Rapport. Learning Gain does not have any significant relationship with any of the curve properties.

Table 2. The table describes the analysis of the fitted curve parameters & properties and their relationships to rapport and learning. a1 and a2 are the coefficients of predictors in the regression based on Equation 12. a, b, RMSE, and R^2 are curve parameters.

<table>
<tr><th colspan="8">Fitted Logarithm Curve (Complete) Parameter Descriptives</th></tr>
<tr><td colspan="2"><i>RMSE</i>
<i>Mean(SD)</i></td><td colspan="2">R^2
<i>Mean(SD)</i></td><td colspan="2"><i>a</i>
<i>Mean(SD)</i></td><td colspan="2"><i>b</i>
<i>Mean(SD)</i></td></tr>
<tr><td colspan="2">0.13(0.04)</td><td colspan="2">0.86(0.1)</td><td colspan="2">0.04(0.01)</td><td colspan="2">0.17(0.05)</td></tr>
<tr><th colspan="8">Pearson's Correlation b/w Curve Parameters and Final Outcomes</th></tr>
<tr><td colspan="4"><i>Partner Rapport</i>
<i>r(p-value)</i></td><td colspan="4"><i>Learning Gain</i>
<i>r(p-value)</i></td></tr>
<tr><td><i>RMSE</i></td><td>R^2</td><td>a</td><td><i>b</i></td><td><i>RMSE</i></td><td>R^2</td><td><i>a</i></td><td><i>b</i></td></tr>
<tr><td>0.4*
(0.04)</td><td>0.27
(0.19)</td><td>0.06
(0.78)</td><td>0.36
(0.07)</td><td>-0.04
(0.85)</td><td>0.07
(0.73)</td><td>0.15
(0.45)</td><td>-0.004
(0.98)</td></tr>
<tr><th colspan="8">Regression Between Combined Curve Parameters and Final Outcomes</th></tr>
<tr><td colspan="2"></td><td colspan="3"><i>Partner Rapport</i></td><td colspan="3"><i>Learning Gain</i></td></tr>
<tr><td colspan="2"><i>Predictors</i></td><td><i>p-value</i></td><td><i>a1</i>
<i>(p-value)</i></td><td><i>a2</i>
<i>(p-value)</i></td><td><i>p-value</i></td><td><i>a1</i>
<i>(p-value)</i></td><td><i>a2</i>
<i>(p-value)</i></td></tr>
<tr><td colspan="2"><i>a and b</i></td><td>0.02*</td><td>-21.73*
(0.03)</td><td>7.66**
(0.005)</td><td>0.42</td><td>10.16
(0.19)</td><td>-2.2
(0.28)</td></tr>
<tr><th colspan="8">Fitted Logarithm Curve Properties' Descriptives</th></tr>
<tr><td></td><td><i>AUC</i></td><td><i>Turn to 50%</i></td><td><i>Value at 50%</i></td><td><i>Turn to Asymptote</i></td><td><i>Value at Asymptote</i></td><td><i>Turn Diff. (Asy. - 50%)</i></td><td><i>Value Diff. (Asy. - 50%)</i></td></tr>
<tr><td>Mean
(SD)</td><td>0.13
(0.04)</td><td>0.096
(0.03)</td><td>0.066
(0.02)</td><td>1.03
(0.85)</td><td>0.16
(0.04)</td><td>0.93
(0.82)</td><td>0.09
(0.03)</td></tr>
<tr><th colspan="8">Pearson's Correlation b/w Curve Properties and Final Outcomes</th></tr>
<tr><td></td><td><i>AUC</i></td><td><i>Turn to 50%</i></td><td><i>Value at 50%</i></td><td><i>Turn to Asymptote</i></td><td><i>Value at Asymptote</i></td><td><i>Turn Diff. (Asy. - 50%)</i></td><td><i>Value Diff. (Asy. - 50%)</i></td></tr>
<tr><td>Partner Rapport</td><td>0.43*
(0.03)</td><td>-0.44*
(0.02)</td><td>0.45*
(0.02)</td><td>-0.38
(0.06)</td><td>0.17
(0.41)</td><td>-0.37
(0.06)</td><td>-0.15
(0.48)</td></tr>
<tr><td>Learning Gain</td><td>-0.05
(0.79)</td><td>-0.11
(0.59)</td><td>-0.22
(0.28)</td><td>-0.14
(0.49)</td><td>-0.01
(0.95)</td><td>-0.14
(0.48)</td><td>0.16
(0.47)</td></tr>
<tr><th colspan="8">Regression Between Combined Curve Properties and Final Outcomes</th></tr>
<tr><td colspan="2"></td><td colspan="3"><i>Partner Rapport</i></td><td colspan="3"><i>Learning Gain</i></td></tr>
<tr><td colspan="2"><i>Predictors</i></td><td><i>p-value</i></td><td><i>a1</i>
<i>(p-value)</i></td><td><i>a2</i>
<i>(p-value)</i></td><td><i>p-value</i></td><td><i>a1</i>
<i>(p-value)</i></td><td><i>a2</i>
<i>(p-value)</i></td></tr>
<tr><td colspan="2"><i>Turn to 50%</i>
<i>&</i>
<i>Value at 50%</i></td><td>0.03*</td><td>-3.5
(0.2)</td><td>5.53
(0.16)</td><td>0.23</td><td>-4.93
(0.1)</td><td>-2.83
(0.18)</td></tr>
<tr><td colspan="2"><i>Turn to Asy.</i>
<i>&</i>
<i>Value at Asy.</i></td><td>0.17</td><td>-0.18
(0.09)</td><td>0.38
(0.87)</td><td>0.74</td><td>-0.59
(0.73)</td><td>-0.06
(0.45)</td></tr>
</table>

We also looked if certain meaningful combinations of these properties can define these final outcomes. We calculated 4 regressions models as per the equation 12, 2 with *Turn to 50%* & *Value at 50%* as $P1$ and $P2$ respectively, and Partner Rapport and Learning Gain as IV; and 2 with *Turn to Asymptote* &

Table 3. The table describes a comparison between early curves and complete curves, with early curve parameters significantly correlated with the complete curve parameters. a and b are the slope and intercept of the fitted logarithmic curves, and RMSE and $R2$ represent the root mean squared error and the r-squared values of the curves.

Early Logarithm Curve Parameter Descriptives							
RMSE *Mean(SD)*		R^2 *Mean(SD)*		*a* *Mean(SD)*		*b* *Mean(SD)*	
0.07(0.02)		0.74(0.16)		0.03(0.01)		-0.02(0.02)	
Pearson's correlations							
Early Curve *&* *Complete Curve* *Parameters*		*Early Classical Metrics* *&* *Final Outcomes*				*Early & Final* *Classical Metrics*	
		Partner Rapport *r(p-value)*		*Learning Gain* *r(p-value)*			
a	*b*	*Early* *Value*	*Early* *Mean*	*Early* *Value*	*Early* *Mean*	*Early Value* *&* *Last Value*	*Early Mean* *&* *Total Mean*
0.49* (0.01)	-0.44* (0.02)	0.32 (0.11)	0.41* (0.03)	-0.13 (0.54)	-0.11 (0.59)	0.57** (0.002)	0.45* (0.02)

Value at Asymptote as $P1$ and $P2$ respectively, and Partner Rapport and Learning Gain as IV. Table 2 shows the results. The joint regression model with *Turn to 50%* and *Value at 50%* as predictors seem to significantly represent Partner Rapport (p-value < 0.05), however, the individual coefficients of these predictors do not show statistical significance in this prediction. Further, there is no other significant relationship between joint predictors and Learning Gain.

This temporal analysis was conducted using fitted curves based on the whole conversation data. We further fitted "Early Logarithmic Curves" just using the first 25 turns of each dyadic session. Table 3 shows the descriptives of the RMSE, the R^2 values, the slope values (a), and the intercept values(b) of the fitted curves. RMSE and R^2 values for these curves represent a good fit. We further calculated direct correlations between the parameters of these early curves and the original curves, as we wanted to see if early behaviors in the conversation could predict the overall trajectories. Table 3 shows the r-value and the p-value of the correlations between the slope (a) and intercept (b) parameters of these curves. We see strong statistically significant correlations between these parameters. We further wanted to see if the values of classical metrics at this early stage can determine their final values. We do see significant correlations between the values in table 3. Moreover, we also assessed if these early values of classical metrics can predict the final Partner Rapport and Learning Gain. We only see (table 3) a strong significant correlation between the early mean value and partner rapport.

Comparing Regression Models with Classical and Temporal Metrics. We also compared whether combining the Temporal and Classical metrics could improve the power of the joint model to predict final outcomes. We did this comparison

Table 4. The table shows a comparison of Regression models with the combination of Classical Metrics and Fitted Curve properties. p-value and Adjusted R^2 represent the parameters of the fitted regression model and the F-statistic p-value is the p-value of the comparison between pairs of regression models with the null hypothesis that the full model does not add any significance to the subset model.

			Predictors						
Classical Metric	*Dependent Variable*	*Regression Params*	*V1*	*V1 + Curve Params.*	*V1 + AUC*	*V1 + Turn to 50%*	*V1 + Val at 50%*	*V1 + Turn to Asy.*	*V1 + Val at Asy*
Last Value (V1)	*Partner Rapport*	*p-value*	0.03*	0.008**	0.09	0.06	0.05	0.08	0.01*
		Adjusted R2	0.14	0.33	0.11	0.15	0.16	0.12	0.24
		F-statistic p-value	-	0.05	0.63	0.27	0.24	0.47	0.05
	Learning Gain	*p-value*	0.85	0.43	0.26	0.84	0.22	0.3	0.9
		Adjusted R2	-0.04	-0.005	0.03	-0.07	0.04	-0.06	-0.07
		F-statistic p-value	-	0.26	0.10	0.58	0.09	0.45	0.67
Mean Value (V1)	*Partner Rapport*	*p-value*	_0.03*_	0.04*	0.07	0.06	0.06	*0.07*	0.01*
		Adjusted R2	0.15	0.21+	0.13+	0.15	0.14	0.13	0.25
		F-statistic p-value	-	0.15	0.48	*0.29*	0.36	0.48	0.04*
	Learning Gain	*p-value*	0.79	0.45	0.32	0.58	0.27	0.5	0.92
		Adjusted R2	-0.04	-0.01	0.02	-0.04	0.03	-0.02	-0.08
		F-statistic p-value	-	0.29	0.14	0.32	0.11	0.26	0.77

by fitting 2 regression models, first with just one Classical Metric, and the second with one Classical and one Temporal metric as predictors. We then compared the Adjusted R^2 and the p-value of the F-test for both models. Table 4 shows the results of this comparison. Combining curve parameters with the Last Value and Mean Value metrics produced a better R^2, however, this increase was not statistically significant as depicted by the p-value of the F-statistic. For the combination of mean and curve parameters as predictors, high multicollinearity was observed, indicating curve parameters are highly correlated with the mean value. The only statistically significant (F-statistic p-value < 0.05) difference between the regressions was found in the combination of Mean Value and Value at Asymptote, with also the bigger model having more Adjusted R^2 value.

4 Discussion

In this work, we analyzed the temporal behaviors of human-human lexical alignment by fitting logarithmic curves to conversational data in a human–human-robot teaching setting. Our analysis compared these temporal metrics with classical alignment measures, i.e. mean value, last-turn values, and total turns to predict partner rapport and learning gain. add more citations in discussion

A common trend observed across the analysis was that none of the alignment metrics, whether classical or temporal, could significantly predict learning gain. We further inspected learning gain distributions which revealed strong departures from normality. The overall distribution and those from Study 1 and Study 2 individually, appeared bimodal and widely dispersed. Such non-normality may partly explain why standard regression and correlation analyses did not capture significant predictive relationships. Also, probable variabilities in pre/post tests across the studies or potential human error in grading may further contribute to this pattern. In the future, we plan to employ mixture models (e.g., Gaussian mixture models) to better account for and explain these bimodal distributions.

For Partner Rapport, classical metrics, e.g. last-turn value, mean alignment, and total turns were all significant predictors. These measures are highly intercorrelated, indicating longer conversations enhance convergence, yielding higher overall alignment values and increased means leading to more rapport. This finding is observed in prior research on alignment in human-robot dialogues [2].

Our curve-fitting approach showed promising results, with very low RMSE and high R^2 values, suggesting that a logarithmic function can capture the overall alignment trends. However, it is important to note that not all alignment trajectories conform perfectly to a logarithmic trend; some curves exhibit dips or fluctuations that our smooth log model does not capture fully. Such deviations could be due to noise inherent in our n-gram-based alignment score computation or the inherently dynamic nature of conversational alignment [3].

Further analysis revealed an intriguing finding with a positive correlation between RMSE and partner rapport, indicating higher RMSE values are associated with better rapport. One interpretation could be that greater variability in alignment (i.e., more “error” from the ideal curve) may indicate dynamic adjustments during conversation, potentially signaling heightened engagement or rapport. However, a deeper qualitative analysis of the actual dips/peaks in the alignment can help in further reflecting on this intuition.

While the individual logarithmic curve parameters, slope, and intercept, did not independently predict rapport, a combined regression model revealed that these parameters significantly determine partner rapport. Specifically, a negative slope (a slower rate of growth in alignment) and a higher intercept (a higher baseline alignment) were strongly associated with partner rapport. This could occur because the parameters might jointly capture variance that neither alone can explain. For instance, baseline alignment could reflect pre-existing rapport factors that are later changed by gradual alignment.

We observed that temporal features such as AUC, Turn to 50%, and the Value at 50%, significantly predict partner rapport. Notably, Turn to 50% showed a negative relationship with rapport, suggesting that quickly achieving a moderate alignment is more indicative of rapport. In contrast, higher AUC and Value at 50% indicate overall stronger alignment, and thus, higher rapport. The measures based on the asymptotic phase did not yield any significant relationship with partner rapport. For 9 out of 26 dyadic sessions, the asymptote was not reached, maybe due to limited time constraints for each session. Even though these asymptotic values did not predict rapport, they still indicate the potential level of alignments that could occur in a dyad and can predict the approximate time it could take for a group to reach these high values. Interventions can be planned in the conversations to accelerate this growth towards the asymptote.

The duration and value change between the 50% point and the asymptote did not emerge as significant predictors of rapport. Further, the combination of features such as Turn to 50% and Value at 50% in a regression model, significantly predicted rapport, though the individual coefficients lost significance. Conversely, models incorporating Turn to Asymptote and Value at Asymptote were not significant. These results suggest that while the early evolution of alignment is critical for rapport, later-stage dynamics may require further exploration.

Notably, parameters from early fitted curves (based on only the first 25 turns) were significantly correlated with the curves derived from full session data. This finding indicates that early interaction dynamics can serve as predictive markers for final partner rapport. This presents an opportunity to design early interventions with a teachable robot that might alter the trajectory of human-human alignment and, consequently, rapport outcomes. From the classical metrics measured at early stages, only the mean showed predictive potential for later values. However, it still is uncertain if this relationship holds uniformly across all conversational turns, as we just computed it at a particular turn.

Finally, our results suggest that, in some cases, augmenting classical metrics with temporal characteristics enhances the predictive power of models for partner rapport. For instance, the combination of mean alignment with Value at Asymptote outperformed a regression model based solely on the mean, highlighting the importance of incorporating temporal dynamics. Although many adjusted R^2 values were high for other combinations, the overall F-statistics were not significant, likely due to the limited sample size. We expect that with a larger dataset, we might be able to get a clearer picture of the distinct contributions of temporal dynamics versus classical measures.

Our future work will extend these analyses by (i) employing mixture models to more accurately characterize bimodal learning gain distributions, (ii) exploring domain-based and semantic alignment measures to complement the lexical metrics, and (iii) investigating problem boundaries and context-specific alignment to determine if these factors better explain learning and rapport outcomes.

Our study has several limitations. First, we are restricted by a small sample size. Further, we only focused on human-human conversations within the human-human-robot teaching context, which may limit the generalizability of our findings. Also, this is an exploratory study, and more experiments are required to strongly signify our results. Further, our alignment metric, based on n-grams, might not capture the full complexity of conversational dynamics. Moreover, our assumption that the entrainment curves follow a logarithmic trend may oversimplify the true underlying dynamics, as real data often exhibit dips or fluctuations that the model smooths over.

5 Conclusion

In this work, we explore the temporal dynamics of human-human lexical alignment in a human-human-robot teaching context. We fit logarithmic curves to visualize these dynamics and use the properties of the fitted curves to predict the final outcomes of learning and rapport. We find that these temporal dynamics are an important predictor of how the alignment evolves and can be used to plan interventions to alter the final outcomes. By integrating dynamic measures of alignment into our analysis, we aim to contribute to a richer understanding of how conversational coordination influences rapport and learning in human–human-robot settings.

Acknowledgments. Left blank for review.

References

1. Argyriou, P., Mohr, C., Kita, S.: Hand matters: Left-hand gestures enhance metaphor explanation. J. Exp. Psychol. Learn. Mem. Cogn. **43**(6), 874–886 (Jun 2017)
2. Asano, Y., Litman, D., Yu, M., Lobczowski, N., Nokes-Malach, T., Kovashka, A., Walker, E.: Comparison of lexical alignment with a teachable robot in human-robot and human-human-robot interactions. arXiv preprint arXiv:2209.11842 (2022)
3. Bonin, F., Looze, C.D., Ghosh, S., Gilmartin, E., Vogel, C., Polychroniou, A., Salamin, H., Vinciarelli, A., Campbell, N.: Investigating fine temporal dynamics of prosodic and lexical accommodation. In: Interspeech (2013), https://api.semanticscholar.org/CorpusID:12027122
4. Doyle, G., Yurovsky, D., Frank, M.C.: A robust framework for estimating linguistic alignment in twitter conversations. In: Proceedings of the 25th international conference on world wide web. pp. 637–648 (2016)
5. Dubuisson Duplessis, G., Clavel, C., Landragin, F.: Automatic measures to characterise verbal alignment in human-agent interaction. In: Jokinen, K., Stede, M., DeVault, D., Louis, A. (eds.) Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. pp. 71–81. Association for Computational Linguistics, Saarbrücken, Germany (Aug 2017). https://doi.org/10.18653/v1/W17-5510, https://aclanthology.org/W17-5510/
6. Dubuisson Duplessis, G., Langlet, C., Clavel, C., Landragin, F.: Towards alignment strategies in human-agent interactions based on measures of lexical repetitions. Language Resources and Evaluation **55**, 353–388 (2021)
7. Dubuisson Duplessis, G., Langlet, C., Clavel, C., Landragin, F.: Towards alignment strategies in human-agent interactions based on measures of lexical repetitions. Language Resources and Evaluation **55**(2), 353–388 (Jun 2021). https://doi.org/10.1007/s10579-021-09532-w, https://doi.org/10.1007/s10579-021-09532-w

8. Duran, N.D., Paige, A., D'Mello, S.K.: Multi-level linguistic alignment in a dynamic collaborative problem-solving task. Cognitive Science **48**(1), e13398 (2024)
9. Friedberg, H., Litman, D., Paletz, S.B.: Lexical entrainment and success in student engineering groups. In: 2012 ieee spoken language technology workshop (slt). pp. 404–409. IEEE (2012)
10. Gratch, J., Wang, N., Gerten, J., Fast, E., Duffy, R.: Creating rapport with virtual agents. In: Pelachaud, C., Martin, J.C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) Intelligent Virtual Agents. pp. 125–138. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
11. Hirschberg, J.B., Nenkova, A., Gravano, A.: High frequency word entrainment in spoken dialogue (2008)
12. Hoegen, R., Aneja, D., McDuff, D., Czerwinski, M.: An end-to-end conversational style matching agent. In: Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents. pp. 111–118 (2019)
13. Lubold, N.: Producing Acoustic-Prosodic Entrainment in a Robotic Learning Companion to Build Learner Rapport. Ph.D. thesis (2018), https://pitt.idm.oclc.org/login?url=https://www.proquest.com/dissertations-theses/producing-acoustic-prosodic-entrainment-robotic/docview/2154434278/se-2, copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-11-01
14. Niederhoffer, K.G., Pennebaker, J.W.: Linguistic style matching in social interaction. Journal of Language and Social Psychology **21**(4), 337–360 (2002)
15. Perner, J.: Theory of mind (1999)
16. Rahimi, Z., Kumar, A., Litman, D.J., Paletz, S., Yu, M.: Entrainment in multiparty spoken dialogues at multiple linguistic levels. In: Interspeech. pp. 1696–1700 (2017)
17. Rahimi, Z., Litman, D.: Entrainment2vec: Embedding entrainment for multiparty dialogues. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 8681–8688 (2020)
18. Reitter, D., Moore, J.D.: Alignment and task success in spoken dialogue. Journal of Memory and Language **76**, 29–46 (2014)
19. Sinha, T., Cassell, J.: We click, we align, we learn: Impact of influence and convergence processes on student learning and rapport building. In: Proceedings of the 1st Workshop on Modeling INTERPERsonal SynchrONy And InfLuence. p. 13–20. INTERPERSONAL '15, Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2823513.2823516, https://doi.org/10.1145/2823513.2823516
20. Srivastava, S., Theune, M., Catala, A.: The role of lexical alignment in human understanding of explanations by conversational agents. In: Proceedings of the 28th International Conference on Intelligent User Interfaces. pp. 423–435 (2023)
21. Srivastava, S., Wentzel, S.D., Catala, A., Theune, M.: Measuring and implementing lexical alignment: A systematic literature review. Computer Speech & Language p. 101731 (2024)
22. Thomas, P., McDuff, D., Czerwinski, M., Craswell, N.: Expressions of style in information seeking conversation with an agent. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 1171–1180 (2020)
23. Tickle-Degnen, L., Rosenthal, R.: The nature of rapport and its nonverbal correlates. Psychological Inquiry **1**(4), 285–293 (1990), http://www.jstor.org/stable/1449345

24. Wang, B., Theune, M., Srivastava, S.: Examining lexical alignment in human-agent conversations with gpt-3.5 and gpt-4 models. In: International Workshop on Chatbot Research and Design. pp. 94–114. Springer (2023)
25. Ward, A., Litman, D.: Dialog convergence and learning. Frontiers in Artificial Intelligence and Applications **158**, 262 (2007)
26. Ward, A., Litman, D.J.: Automatically measuring lexical and acoustic/prosodic convergence in tutorial dialog corpora (2007)
27. Yu, M., Litman, D., Ma, S., Wu, J.: A neural network-based linguistic similarity measure for entrainment in conversations. arXiv preprint arXiv:2109.01924 (2021)