# User Affect and No-Match Dialogue Scenarios: An Analysis of Facial Expression

Joseph B. Wiggins[1], Mayank Kulkarni[1], Wookhee Min[2], Kristy Elizabeth Boyer[1], Bradford Mott[2], Eric Wiebe[3], and James Lester[2]

[1]Department of Computer & Information Science & Engineering, University of Florida, Gainesville, FL 32601

[2]Center for Educational Informatics, North Carolina State University, Raleigh, NC 27695

[3]STEM Education, North Carolina State University, Raleigh, NC 27695

[1]{jbwiggi3, mayankk91, keboyer}@ufl.edu, [2]{wmin, bwmott, lester}@ncsu.edu, [3]wiebe@ncsu.edu

## ABSTRACT

Recent years have seen significant advances in natural language dialogue management and a growing recognition that multimodality can inform dialogue policies. A key dialogue policy problem is presented by 'no-match' scenarios, in which the dialogue system receives a user utterance for which no matching response is found. This paper reports on a study of the 'no-match' problem in the context of a dialogue agent embedded within a game-based learning environment. We investigate how users' facial expressions exhibited in response to the agent's no-match utterances predict the users' opinion of the agent after the interaction has completed. The results indicate that models incorporating users' facial expressions following no-match utterances are highly predictive of user opinion and significantly outperform baseline models. This work represents a key step toward affect-informed dialogue systems whose policies are informed by users' affective expression.

## CCS CONCEPTS

• **Human-centered computing** → Empirical studies in HCI • **Human-centered computing** → Natural language interfaces

## KEYWORDS

Dialogue agents, facial expression, no-match dialogue policy

**ACM Reference format:**

J. B. Wiggins, M. Kulkarni, W. Min, B. Mott, K. E. Boyer, E. Wiebe, and J. Lester. 2018. User Affect and No-Match Dialogue Scenarios: An Analysis of Facial Expression. In International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction (MA3HMI'18), October 16, 2018, Boulder, CO, USA. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3279972.3279979



MA3HMI'18, October 16, 2018, Boulder, CO, USA

ACM ISBN 978-1-4503-6076-0/18/10…$15.00
https://doi.org/10.1145/3279972.3279979

## 1 INTRODUCTION

Designing dialogue systems has long been a focus of the ICMI community. For example, multimodal considerations are central in the design of human-robot dialogue interactions [26] and for dialogue systems that engage in interviews [8], training [35], and interactions with children [42]. Dialogue *policies* determine what conversational moves a system makes at a given time. Dialogue policies are typically handcrafted, machine-learned, or some combination thereof. The design and evaluation of dialogue policies has made great strides in recent years, with the increasing use of machine learning and data-driven approaches [28] and crowdsourcing [37], along with an increasing appreciation for affective phenomena such as engagement and frustration [14, 15, 18, 19].

This growing body of work, along with seminal human dialogue research [21, 22, 43], suggests that users' multimodal expressions including prosody, gesture, gaze, and facial expression, can provide deep insight into user experience, and therefore can and should inform dialogue policies. However, the ways in which multimodal user expressions can best be used to inform adaptive dialogue policies is an open question. This paper investigates that question with a specific focus on the inevitable, yet under-researched, 'no-match' scenario, in which the user encounters a limitation of the dialogue system's ability to interpret or respond [7, 36]. A common dialogue policy for no-match scenarios has traditionally been to deliver an utterance such as, "I'm sorry, I didn't understand that" [24] or "I'm not sure I can talk about that" [36]. However, conversational user interface design requires careful awareness of the impact such utterances may have on users, such as undermining their belief that the dialogue agent is knowledgeable or helpful [16]. Multimodal features such as user facial expressions provide an excellent source of information about the user's state, and can be used to create adaptive dialogue policies for no-match scenarios that take the individual user experience into account.

This paper investigates how users' facial expressions exhibited in response to the agent's no-match utterances predict the users' opinion of the agent after the interaction has completed.

In an analysis of user interactions with a dialogue agent embedded in a game-based learning environment, we find that the *number of no-match utterances* a user received is slightly predictive of that user's opinion of the agent, and the *sentiment* expressed through natural language by users in response to those utterances is not at all predictive. However, the user's *facial expressions* within three, five, and seven seconds after receiving no-match utterances are highly predictive of their opinion of the agent.

This paper makes the following contributions: 1) it explores how no-match dialogue system utterances were received by users; 2) it explores a nonverbal channel, namely, facial expression reactions to dialogue and their relationship to users' opinion of the system they are conversing with; and 3) it explores the predictive power of facial expression across different windows of time, and the potential of mixed time-window approaches. This work represents a key step toward individually adaptive dialogue systems whose policies are informed by affective expressions of users.

## 2 RELATED WORK

We investigate affect-informed no-match dialogue moves in the context of a game-based learning environment that supports users in learning microbiology. Because the dialogue agent considered here has an implicitly pedagogical purpose (*i.e.*, to support learners as they complete a learning task), we discuss related work on pedagogical agents before turning to related work on no-match dialogue policies.

### 2.1 Pedagogical Agents

Supporting learning is a particularly promising application area for dialogue agents, and agents that support learning are often referred to as *pedagogical agents*. Pedagogical agent research has been conducted in domains including the circulatory system [25], helping with homework [34], negotiation skills [20] and training for job interviews [11]. Pedagogical agents can help students reflect on their current knowledge and determine areas where further learning is needed [2] and manage their frustration during learning [9, 10]. The dialogue design of pedagogical agents can have a substantial impact on students' learning outcomes [27].

There are still many open questions regarding how students perceive and interact with pedagogical agents. While pedagogical agents have shown great potential for positive impact, this impact is not achieved with all students [45] and can be inconsistent across groups [41]. Empirical investigations have only begun to answer these questions, with results showing that the most suitable representation of an agent differs by domain [39], that designers must be aware of sociocultural concerns when designing multimodal interactions with agents [38], and that introducing an agent may distract students from learning goals if nuances of human memory capacity are not carefully considered [31].

Work has begun to explore how user multimodal expressions may be important in understanding critical points in dialogue. In particular, user facial expressions during dialogue interactions can be highly informative for learning. For example, facial expression has been used to model student certainty when answering a question [5]. Facial expressions play a role in communication even if speakers are not physically co-located [16]. Multimodal measures have been shown to have significant relationships with important outcomes including engagement and frustration [15, 16].

### 2.2 No-Match Dialogue Policies

Pitterman and colleagues formally defined a no-match scenario as a situation in which rules or classes of linguistic analysis are unable to match the content of the user's input or turn [36]. No-match responses can be regarded as indicators of how the dialogue interactions are unfolding. Clemens and Hempel [7] consider the number of no-matches as an indication that the user might need more explanation, suggesting that systems should redefine its dialogue strategy by reducing the possible input state space to obtain better match accuracy. Pitterman and colleagues also explored using no-match responses as an assessment of how effective a dialogue strategy is for a user [36]. They consider no-match responses as an indicator of how experienced a user is, with typically fewer no-match corresponding to a more experienced user. This allows dialogue designers to define different strategies based on their perceived user experience without treating every user of the system as a novice.

No-match responses are also generally thought of as a phenomena to avoid. Varges and colleagues [46] explored automatic constraint relaxation strategies and found users preferred some sort of suggestion instead of a simple no-match response. Jin and colleagues [23] investigated whether it was more useful to choose a no-match response over a possibly incorrect response, and found that choosing possibly incorrect responses reduces the no-match responses significantly while only slightly increasing the actual incorrect responses.

None of the previous work has taken user affect into account. In educational domains such as the one that is the focus of this work, no-match scenarios could be important moments for students to clarify their thoughts. However, probing them to do so could introduce frustration. During early piloting of the dialogue agent presented in this paper, researchers noted students' frustration around no-match scenarios. This paper explores their reactions to no-match agent utterances, with the intention of uncovering how those utterances affect a student's opinion of the dialogue agent they are interacting with. The remainder of this paper is structured as follows: we discuss the dialogue agent that was used in this work, describe our methodology, the multimodal data that was collected, and the analysis, and conclude with a discussion of implications for multimodal dialogue research.

## 3 DIALOGUE AGENT

This section describes the dialogue agent we built to support users within a game-based learning environment for microbiology education. The premise of the game is that an outbreak on a remote research station is underway. The student, playing the role of a medical field agent, is given the responsibility of identifying the disease and its source. The student explores the environment from a first-person perspective (Figure 1). The student's objective

is to diagnose the outbreak. To do so, the student must gather evidence from virtual characters and explore in-game educational resources, such as books and quizzes, and test hypotheses using virtual lab equipment. Students characteristically experience a breadth of emotion during this challenging task, presenting many opportunities to provide support through a dialogue agent.

**Figure 1: The first-person perspective of a user in the game-based learning environment.**

The Alisha dialogue agent is presented as part of the game world, and she introduces herself as virtual field agent in training, working for the Center of Disease Control (CDC). Her virtual-agent-in-training status is designed to provide a narrative device through which she can smoothly guide dialogues back to on-task conversation. The interface (Figure 2) allows students to exchange textual messages with Alisha, who was accessible to the student any time except in two contexts: while engaged in menu-based interaction with other game characters, and while interacting with books or embedded assessments.

## 3.1 Dialogue Manager

Upon receiving a user utterance, the dialogue manager first attempts to check for an exact match with a predefined set of iteratively refined question answers (Figure 3). If the system is unable to find a match for an utterance to a question, it parses the given utterance for game objects such as people, items, or books. If a game object is identified, a dependency parse is performed to obtain the associated verb, if any, to understand the question's intention with respect to the game object. The system then attempts to respond through a rule-based natural language generator. If none of the rules are matched, the system removes the stop words from the utterance and tries to perform a partial match on the predefined question answer set. If fewer than two words are matched and the utterance does not contain a word for a game object, the utterance-question pair is not considered as a partial match. If there is no partial match obtained, the system then attempts to find the most similar sentence vector from the question answer set to the given utterance's vector representation obtained by using Google News word vectors.[2] The sentence vector representation is obtained by averaging the word vector representations of the unique words in a sentence. The similarity between two sentences is measured by the cosine similarity of those two sentences' vector representations, a standard technique in sentence comparison. The threshold for considering the sentence similar is 0.85. If there is no similar sentence then we apply a threshold of 0.6 with the goal of matching the high-level topic of the question and generating a clarification statement. These thresholds were determined over a series of user interactions with the system prior to conducting the study presented here. Finally, if no topic is found, the system sends a no-match response for the given utterance (Table 1).

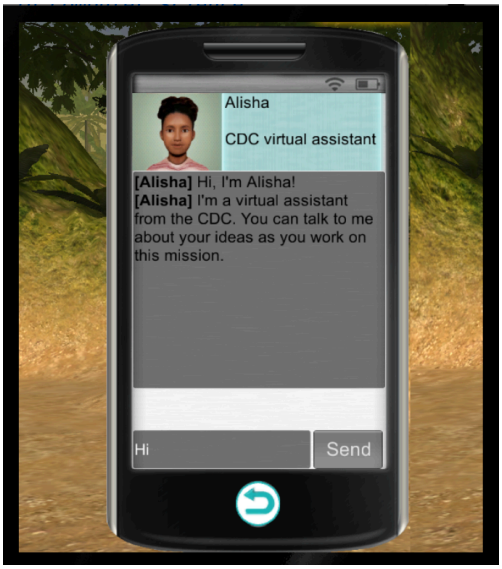



**Figure 2: The dialogue agent interface.**

Alisha's no-match utterances take a variety of forms. Each holds the goal of informing the user that, given the current information, the system is unsure of an appropriate response. The no-match utterances used in the system can be seen in Table 1.

**Table 1: Examples of No-Match Utterances**

| |
|---|
| I didn't quite get what you said. |
| If you're looking for more information search for books or posters. |
| I don't understand, talking to people on the island might help. |
| I didn't understand, could you try rephrasing your question. |
| I'm sorry I don't understand. |

[2] https://code.google.com/archive/p/word2vec/

### 3.2 Student Dialogue Responses to No-Match Utterances

As students progressed through the learning environment and interacted with the dialogue agent, some students made utterances that were beyond the current capabilities of the system. In these cases a no-match utterance was given. A scenario of a no-match dialogue is illustrated in Table 2, in which a student said something the agent could not understand, the no-match utterance was given, and then the student repaired what they had originally stated in a form the agent could respond to.

**Table 2: Dialogue Excerpt of the Resolution of a No-Match**

| | |
|---|---|
| S12: | where is bryce |
| Agent: | Bryce is in the house near the kitchen. |
| S12: | where is the kitched |
| Agent: | I don't understand, talking to people on the island might help. |
| S12: | where is the kitchen |
| Agent: | The kitchen is towards the left side of the campsite. |

However, no-match responses may also lead to frustration from the student, or feelings that the dialogue agent is not helpful. An excerpt highlighting this scenario can be seen in Table 3. In it, the user made a statement the agent could not match, the agent responded with a no-match and then the student insulted the agent, calling her an "idiot." This situation is an instance of students expressing frustration verbally through dialogue, but as we will see in the next section, many students simply choose not to respond to no-match agent utterances. In these cases, facial expression can provide valuable insight into the user's experience with the dialogue agent.

**Table 3: Dialogue Excerpt of Student Frustration Resulting from a No-Match**

| | |
|---|---|
| Agent: | Speaking to the virus specialist, Ford, might give us some clues on how to proceed. |
| S19: | Where is he or she? |
| Agent: | I didn't quite get what you said. |
| S19: | idiot |

## 4 METHODOLOGY

The study reported here was conducted in spring 2018. This section describes the participants, procedure, and instruments used for data collection.

### 4.1 Participants

The participants for this study were recruited from an introductory programming course for computer science majors, taught at a large university during the spring 2018 term. The participants received extra credit in their course to participate in the study. The study was approved by the university's human subjects review board and informed consent was received from each participant, after an explanation of what data were being collected and the participants' ability to withdraw from the study at any time.

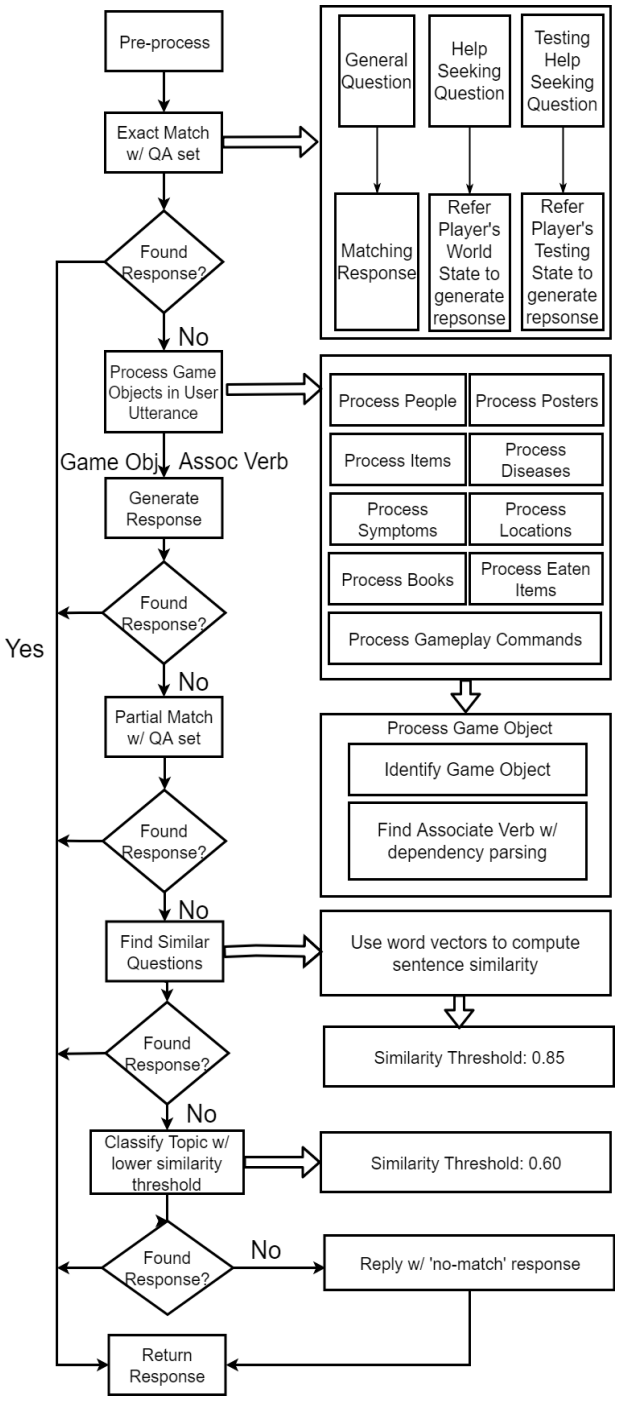



**Figure 3: The flow of a user utterance through the response system, bottoming out with a no-match.**

Each participant's dialogue with the agent unfolded differently, and in the analyses reported here we only considered students who received one or more no-match utterances. Out of 34 participants in the study, 21 students encountered a no-match response in their session. Only 19 of those had intact facial expression tracking during the response, for the reasons explained in described in section 5. The demographics of these students are shown in Table 4.

**Table 4: Participants**

| Feature | Details |
|---|---|
| Age (Average; Std. dev) | 18.9; 0.62 |
| Gender | 6 Female, 13 Male |
| Language Most Comfortable Reading | 18 English, 1 Spanish |
| Video Game Experience (*How frequently do you play video games?*) | 4 Very Frequently<br>3 Frequently<br>8 Occasionally<br>4 Rarely |

## 4.2 Procedure

Participants interacted with the dialogue agent while playing the game-based learning environment and then completed a post-survey (Section 4.3). Participants played for one hour or until they solved the in-game mystery. Before they began playing, the study administrators encouraged the participants to consult the virtual agent whenever they had a question.

## 4.3 Instruments

Data was collected before, during, and after students' interactions with the learning environment. Prior to interaction, we administered validated surveys to measure growth mindset (the belief that intelligence can be increased with effort) [12] and self-efficacy in three domains: science [12], enlisting social resources [3], and self-regulated learning (e.g., actively choosing study strategies) [3]. Prior to interaction, participants also completed a pre-test consisting of multiple-choice questions on microbiology. During gameplay, all student actions and facial expression analyses (Section 5) were logged. After an hour of gameplay, surveys were used to assess self-reported engagement with the game [33], student experience with the virtual learning companion, and overall student affective experience [17]. A content knowledge post-test (identical to the pre-test) was administered upon the completion of the study. Finally, demographic information was collected after all other instruments had been completed.

The focus of this work is on the student's opinion of the dialogue agent. After interacting with the agent, the students were asked to reflect on their interactions with the virtual agent and then rate each item shown in Table 5 on a 5-point Likert scale: "Strongly Disagree", "Disagree", "Neither Agree Nor Disagree", "Agree", or "Strongly Agree." To convert these ratings into a numeric value for model building, each answer was converted to a 1-5 scale (1 being "Strongly Disagree" and 5 being "Strongly Agree") and then averaged to represent the opinion that the student held for the agent. The items "Alisha interrupted my work." and "Alisha frustrated me." were reverse coded due to the negative sentiment. The agent opinion distribution across participants can be seen in Figure 4.

**Table 5: Survey Items to Assess Student's Opinion of Dialogue Agent (* Indicates Reverse Coding)**

| Survey Item: | Average: |
|---|---|
| Alisha was knowledgeable. | 3.26 / 5 |
| Alisha's input was helpful. | 3.42 / 5 |
| I felt encouraged by Alisha. | 3.00 / 5 |
| Alisha responded to me promptly. | 3.79 / 5 |
| Alisha interrupted my work.* | 2.89 / 5 |
| Alisha encouraged me to be part of the conversation. | 2.89 / 5 |
| I was more motivated working with Alisha than I would have been on my own. | 3.26 / 5 |
| Alisha said useful things to me. | 3.79 / 5 |
| Alisha helped me to concentrate. | 2.95 / 5 |
| I paid attention to what Alisha was saying. | 3.84 / 5 |
| I could understand what Alisha was saying. | 4.16 / 5 |
| Alisha said things at sensible times. | 3.32 / 5 |
| I would like to talk with Alisha again. | 3.00 / 5 |
| I would like to play this game again. | 3.26 / 5 |
| Alisha frustrated me.* | 3.42 / 5 |

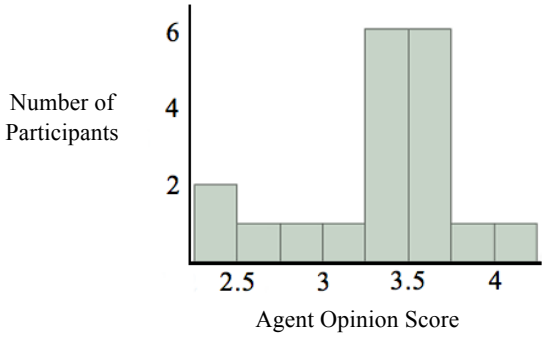


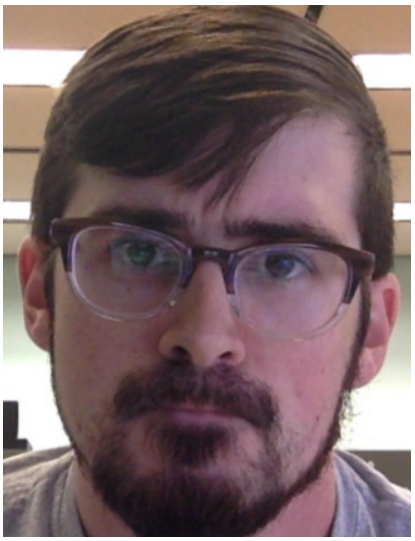

**Figure 4: Distribution of self-reported agent opinion score.**

# 5 MULTIMODAL DATA

As students interacted with the dialogue agent through the game-based learning environment, dialogue and game logs were saved. Additionally, user facial expressions were tracked in real time using Affectiva's AFFDEX SDK [30] and logged at one-second intervals. These facial action units are possible muscle movements in the face, each of which can be activated to varying degrees. The AFFDEX SDK also provides several facial expression estimates, including composite measurements such as Joy, Sadness, Anger, and also what are referred to as Facial Action Units that correspond to the Facial Action Coding System

(FACS) [11]. Each measurement is represented on a scale from 0 (absent) to 100 (present) [11]. An example can be seen in Figure 5.

During student sessions, facial expression recording sometimes lapsed momentarily. The student's face may have become partially occluded; for example, some students would place their hand over their face or cover their face with their shirt. If a lapse took place at any time during the window following a no-match event, this event was discarded from analysis.



| Composite Measurements: | Facial Action Units: |
|---|---|
| Joy | AU17 – Lip Suck |
| <0.0001 | 93.3109 |
| Sadness | AU28 – Eye Closure |
| 98.2991 | <0.0001 |
| Anger | AU43 – Chin Raiser |
| 12.0215 | 80.8802 |

**Figure 5: Example AFFDEX SDK results for a facial expression.**

## 6 ANALYSIS

We analyzed the data to investigate two hypotheses. First, we hypothesized that the number of no-match utterances a user received would have a significant negative correlation with that user's opinion of the agent. Second, based on prior observations that users respond differently to agent no-match utterances, both in terms of what they choose to say verbally and how they express affect nonverbally on the face, we hypothesized that these differing reactions would be predictive of the users' opinion of the agent.

### 6.1 Baseline Model

First, we explored the hypothesis that the frequency of agent no-match utterances is correlated with the user's post-hoc agent opinion. We built a stepwise linear regression model using no-match frequency features to predict the agent opinion, as discussed in Section 4.3. Each row of data represented the percent of agent utterances a user received that were no-match responses, across the 19 students that are the focus of these models. Leave-one-out cross-validated $R^2$ ($LOOCVR^2$) was used as the stopping feature in the stepwise linear regression and reported in the model. Therefore, the model was built using 18 folds, and the predictive accuracy of the model was evaluated using the held-out student in each fold. The results show that the percent of no-match responses was not a significantly predictive feature, as shown in Table 6. We also built a model with absolute frequency of no-match utterances rather than relative frequency, and found an even weaker relationship. That predictor accounted for very little of the variance in students' opinion of the agent ($R^2 = 0.08$).

**Table 6: Baseline Model with Leave-One-Out-Cross-Validated $R^2$ ($LOOCVR^2$)**

| Baseline Model: $R^2 = 0.0809$, $LOOCVR^2 = -0.176$ | | |
|---|---|---|
| **Feature** | **β** | **p-value** |
| Percent No-Match | -1.3967 | 0.2380 |
| Intercept | 3.5336 | <.0001 |

Next, we investigated the hypothesis that we could predict users' opinion of the dialogue agent with the sentiment those users expressed verbally after no-match utterances. Using the Stanford Deeply Moving sentiment analysis toolkit [44], we tagged all student responses to no-match utterances. This process produced probabilities that the utterance fits into each of five tags: *very negative, negative, neutral, positive,* and *very positive*. To represent the sentiment each user expressed, we constructed two different types of features for each of the five tags, resulting in ten potential sentiment features per student. The two feature sets were: 1) *calculating the average sentiment expressed to the agent after a student had received a no-match response*, and 2) *using the highest sentiment rating across any response*. These features capture the general sentiment expressed to the agent and the extremes.

Once again, we constructed a stepwise linear regression model using no-match frequency features to attempt to predict the agent opinion. Each row of data represented the features noted above across the 19 students. Leave-one-out cross-validated $R^2$ was once again used as the stopping feature in our stepwise linear regression, over 18 folds, and we evaluated the predictive accuracy of the model using the excluded student in each fold. None of the sentiment features were a significant predictor of agent opinion. We therefore treat the model shown in Table 6 using frequency of no-match utterances as a baseline, and proceed to investigate whether facial expression can significantly improve predictive power for users' opinion of the agent.

### 6.2 Facial Expression Models

To investigate users' facial expression in response to agent no-match utterances, we considered the automatically tracked and

labeled facial expression data within a time window immediately following no-match agent utterances. We experimented with three different time windows: 3 seconds, 5 seconds, and 7 seconds after the no-match utterance was delivered. Estimates of facial expression were calculated and logged every second, with a range of 0 to 100 as previously described. These features were then averaged across the time window in consideration, and then the average facial expression window values were themselves averaged across each student, producing one row of data per student in the dataset. We provided those facial expression values as predictors to the same linear regression framework described previously to predict the user's opinion of the agent. Stepwise linear regression was used to build the models shown in Table 7 and 8. The models add features which result in the highest leave-one-out cross-validated $R^2$ for predicting user opinion of the agent. The 'Intercept' feature represents the rating that would be expected if all other features were zero.

**Table 7: Facial Expression Models**

| Feature | β | p-value |
|---|---|---|
| Three Second Window: $R^2 = 0.4571$, $LOOCVR^2 = 0.1687$ | | |
| Anger | -15.2426 | 0.0277 |
| Chin Raise | 0.026 | 0.0495 |
| Lip Suck | -0.031 | 0.0232 |
| Intercept | 3.4574 | <.0001 |
| Five Second Window: $R^2 = 0.4610$, $LOOCVR^2 = 0.3025$ | | |
| Anger | -12.6874 | 0.0363 |
| Lip Press | 0.02353 | 0.0993 |
| Lip Suck | -0.04249 | 0.0220 |
| Eye Closure | 0.04064 | 0.1422 |
| Intercept | 3.4309 | <.0001 |
| Seven Second Window: $R^2 = 0.1587$, $LOOCVR^2 = 0.0277$ | | |
| Anger | -11.3751 | 0.0911 |
| Intercept | 3.4170 | <.0001 |

As shown in Table 7, the facial expression models significantly outperform the baseline model. For the three-second time window, four facial expression features explain 16% of the variance in user agent opinion based on leave-one-out cross-validated $R^2$. Examining a five-second window this performance improves to 30% of variance, and then declines considerably to only 2% of variance explained in the seven-second window. We next investigated whether a mixture of features across different time windows would further improve predictive power, as facial expressions have different temporal profiles, with some occurring and fading rapidly while others persist. Indeed, the mixed time window facial expression model (Table 8) explains 64% of variance under a leave-one-out cross-validated $R^2$ framework (Table 9). Models which had access to all the features mentioned in this results section (frequency, sentiment, and facial expression) were also built, but only facial expression features were selected.

**Table 8: Mixed Time Window Facial Expression Models**

| Feature | β | p-value |
|---|---|---|
| Mixed Time Window: $R^2 = 0.8704$, $LOOCVR^2 = 0.6384$ | | |
| Sadness – 3 Seconds | -11.3079 | 0.0001 |
| Chin Raise – 3 Seconds | 0.03633 | 0.0009 |
| Lip Suck – 3 Seconds | -0.4242 | <.0001 |
| Eye Closure – 3 Seconds | -0.08242 | 0.0138 |
| Sadness – 5 Seconds | 10.5741 | 0.0001 |
| Anger – 7 Seconds | -10.8979 | 0.0045 |
| Intercept | 3.5261 | <.0001 |

**Table 9: Model Comparison**

| Model | $R^2$ | $LOOCVR^2$ |
|---|---|---|
| Baseline Model | 0.0809 | -0.176 |
| Three Seconds | 0.4571 | 0.1687 |
| Five Seconds | 0.4610 | 0.3025 |
| Seven Seconds | 0.1587 | 0.0277 |
| Mixed Time Window | **0.8704** | **0.6384** |

## 7 DISCUSSION

We have presented models to predict user opinion of a dialogue agent embedded within a game-based learning environment. These models were based on the frequency of no-match utterances the users received, the sentiment the users expressed verbally in response to the no-match utterances, and the facial expressions they displayed during 3-, 5-, and 7-second windows after the no-match utterance was delivered. The results show that facial expression features are significantly more predictive of user opinion of the dialogue agent than any of the other features, and that a combination of different time windows of facial expression is most predictive.

### 7.1 Facial Expression Predictors of User Opinion

The models using facial expression as predictors were able to outperform the baseline models substantially. Sentiment analysis did not provide any predictive features, though one might expect to see a strong relationship between the valence of the words said to the dialogue agent and the opinion the student forms of that agent. However, as described by Clavel and Callejas [6], traditional sentiment analysis does not account for the multimodal and contextual nature of human-agent interaction. In our domain, the student utterances were usually short, and this may have presented additional challenges to the sentiment analysis. This result in combination with the significant facial expression findings highlight the possibility that what remains *unsaid* in dialogue with an agent may be even more important than what is said.

Previous work has experimented with various time windows in which to consider facial expression activity, finding different

windows important for different affective phenomenon [4]. Other work has used models that account for the sequential nature of facial expressions, such as Long Short-Term Memory (LSTM) models [47]. In our models, we saw strong predictive power from several of the facial expression time windows, but particularly strong predictive power from the model that could sample from all of the time windows. This result speaks to the temporal nature of facial expressions, with some features being important over particular windows of time and not in others.

Several of the facial expression features that were selected by the models have emerged as important within prior work. First, we discuss the specific action units selected by the mixed time window model, and then we discuss the composite features. The action units that were identified as important features following agent no-match utterances were: *lip suck* (AU28), *eye closure* (AU43), and *chin raiser* (AU17). Lip suck (AU28) had a significant negative correlation with the user's rating of the agent in this work. Previous work has pointed to lip suck being negatively associated with presence in game-based learning [40] and there has also been evidence linking it to amusement [29]. This body of work seems to align with our findings, with students potentially feeling that the illusion of the dialogue agent's intelligence is broken or its inability funny in the view of the student. Eye closure (AU43) also had a significant negative association with the user's impression of the dialogue agent in our mixed time window model. Sawyer and colleagues [40] also found this feature to be important in game-based learning environments, and negatively associated with learning gain. Chin raiser (AU17) was the only action unit that was positively associated with agent opinion in our mixed time window model. Chin raiser is most famously related to sadness [13]. Namba and colleagues [32], however, found that chin raiser was of only present in posed negative expressions, not naturally occurring ones. There has also been some weak evidence that it has a positive relationship with amusement [29].

The two composite facial expressions that were selected by the models are anger and sadness. Anger, which was associated with lower opinion of the dialogue agent, has previously been associated with frustration [18] and confusion during tutoring [9]. The relationship with these affective states may be the reason for decreased user opinion of the agent. The other feature, sadness, seems to have a more complicated relationship with user opinion of the dialogue agent. In our mixed time window model, when seen at the three-second time window, it had a significant negative correlation with agent opinion, but then becomes positive at the five-second interval. In previous work, elements of sadness have been related with frustration [18] and confusion [9], which compliments folk knowledge on sadness. However, sadness has also been seen to have a positive relationship to a student's feelings of presence in a game-based learning environment [40] and features that compose sadness have also been found to be related to finding sessions worthwhile [18]. The dynamic nature of this feature highlights the importance of considering progressions of expression, as meaning of some affective expressions may vary over different time windows.

## 7.2 Implications for Designing Dialogue Systems

The models revealed relationships between facial expression reactions to no-match utterances and agent opinion that could not be modeled with either utterance frequency or sentiment features. This result speaks to the importance of multimodal expressions during dialogue with agents. Dialogue system developers should consider facial expressions and other nonverbal displays as contextualized within dialogue. It is also important to consider the reasons that we see these reactions to 'no-matches' having strong relationships to the agent opinion. What we are observing is individual differences in the reception of no-match utterances, despite the same threshold for delivering those no-match utterances used by the system for all users. It seems that when limitations in the dialogue system's capabilities are revealed to the users, their response is varied, with some reacting more negatively than others. For users who are particularly negatively impacted, we need to consider designing new dialogue policies that are adaptively deployed. Alternate policies could potentially focus on building common ground or addressing the affective dimensions of the conversation by focusing on rapport building. A higher threshold for making a "best guess" at an appropriate response could also be considered.

# 8 CONCLUSION

This paper has explored the ubiquitous, yet under-studied, *no-match* dialogue policy. We investigated the hypotheses that the number of no-match utterances, user sentiment in response to those utterances, and facial expression immediately following no-match utterances, would be predictive of user opinion of the dialogue agent as measured after interaction. The results show that only facial expression features were predictive of this outcome, and strongly so. Additionally, it is clear that in some contexts, users may not express their negative experience with an agent verbally, but do so through facial expression. Moreover, we explored three potential time windows (three, five, and seven seconds after a no-match utterance) and found that facial expression features from each of those time windows served as important predictors of the user's opinion of the agent.

In future work, it will be critical to consider how to utilize multimodal features when dynamically adjusting a dialogue policy's thresholds, especially on larger groupings of diverse users. This type of adaptive policy, along with other types of feedback that are supported by multimodal data streams, could significantly improve user experience with dialogue agents.

# ACKNOWLEDGEMENTS

This work is supported by the National Science Foundation through grant IIS-1409639 and DRL-1721160. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

# REFERENCES

[1] S. Attardo, J. Eisterhold, J. Hay, and I. Poggi. 2003. Multimodal markers of

irony and sarcasm. *Humor*, 16, 2 (2003), 243-260.
[2] R. Azevedo, A. Johnson, A. Chauncey, and C. Burkett. 2010. Self-regulated learning with MetaTutor: Advancing the science of learning with MetaCognitive tools. In *New Science of Learning. Springer*, New York, NY, 225-247.
[3] A. Bandura. 2006. Guide for constructing self-efficacy scales. *Self-efficacy Beliefs of Adolescents*, *5*(1), 307-337.
[4] N. Bosch, S. D'Mello, R. Baker, J. Ocumpaugh, V. Shute, M. Ventura, L. Wang, and W. Zhao. 2015. Automatic detection of learning-centered affective states in the wild. *In Proceedings of the 20th International Conference on Intelligent User Interfaces*. ACM, Atlanta, Georgia, 379-388.
[5] A. Bourai, T. Baltrušaitis, and L. P. Morency. 2017. Automatically predicting human knowledgeability through non-verbal cues. *In Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, Glasgow, Scotland, 60-67.
[6] C. Clavel and Z. Callejas. 2016. Sentiment analysis: from opinion mining to human-agent interaction. *IEEE Transactions on Affective Computing*, 7, 1 (2016), 74-93.
[7] C. Clemens and T. Hempel. 2008. Automatic User Classification for Speech Dialog Systems. *In Usability of Speech Dialog Systems*. Springer, Berlin, Heidelberg, 67-80.
[8] K. Cofino, V. Ramanarayanan, P. Lange, D. Pautler, D. Suendermann-Oeft, and K. Evanini. 2017. A modular, multimodal open-source virtual interviewer dialog agent. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, Glasgow, Scotland, 520-521.
[9] S. K. D'Mello, S. D. Craig, and A. C. Graesser. 2009. Multimethod assessment of affective experience and expression during deep learning. *International Journal of Learning Technology*, 4(3-4), 165-187.
[10] S. K. D'Mello, and A. Graesser. 2012. AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems*, 2(4), 23.
[11] I. Damian, T. Baur, B. Lugrin, P. Gebhard, G. Mehlmann, and E. André. 2015. Games are better than books: In-situ comparison of an interactive job interview game with conventional training. *In Proceedings of the 17th International Conference on Artificial Intelligence in Education*. Springer, 84-94.
[12] C. S. Dweck. 1999. Self-theories: Their role in motivation, personality, and development. *Essays in Social Psychology*. Psychology Press: New York, NY, US.
[13] P. Ekman and W. Friesen. 1978. The facial action coding system (FACS): a technique for the measurement of facial action Vol. *Consulting Psychologists*. Palo Alto, CA.
[14] M. Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication*, 51(10), 832-844.
[15] A. Fandrianto, and M. Eskenazi. 2012. Prosodic entrainment in an information-driven dialog system. *In Proceedings of the 13th Annual Conference of the International Speech Communication Association*. Portland, Oregon, USA. 342-345.
[16] Google. In conversation, there are no errors. (October 2017). Retrieved from https://developers.google.com/actions/design/conversation-repair
[17] S. G. Hart and L. E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *In Advances in Psychology* (Vol. 52, pp. 139-183). North-Holland.
[18] J. F. Grafsgaard, J. B. Wiggins, K. E. Boyer, E. N. Wiebe, and J. Lester. 2013. Automatically recognizing facial expression: Predicting engagement and frustration. *In Proceedings of the 6th International Conference on Educational Data Mining*. Memphis, Tennessee, USA, 43-50.
[19] J. F. Grafsgaard, J. B. Wiggins, A. K. Vail, K. E. Boyer, E. N. Wiebe, and J. C. Lester. 2014. The additive value of multimodal features for predicting engagement, frustration, and learning during tutoring. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, Istanbul, Turkey, 43-49.
[20] J. Gratch, D. DeVault, and G. Lucas. 2016. The benefits of virtual humans for teaching negotiation. *In Proceedings of the 16th International Conference on Intelligent Virtual Agents*. Springer, Los Angeles, CA, USA, 283-294.
[21] B. J. Grosz. 2018. Smart Enough to Talk With Us? Foundations and Challenges for Dialogue Capable AI Systems. *Computational Linguistics*, 44(1), 1-15.
[22] B. J. Grosz, and C. L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3), 175-204.
[23] L. Jin, M. White, E. Jaffe, L. Zimmerman and D. Danforth. 2017. Combining CNNs and Pattern Matching for Question Interpretation in a Virtual Patient Dialogue System. *In Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. 11-21.
[24] D. Jurafsky, and J. H. Martin. 2014. Speech and Language Processing (Vol. 3). Pearson, London.
[25] S. Lallé, N. V. Mudrick, M. Taub, J. F. Grafsgaard, C. Conati, and R. Azevedo. 2016. Impact of individual differences on affective reactions to pedagogical agents scaffolding. *In Proceedings of the 16th International Conference on Intelligent Virtual Agents*. Springer, Los Angeles, CA, USA, 269-282.
[26] I. Leite, A. Pereira, A. Funkhouser, B. Li, and J. F. Lehman. 2016. Semi-situated learning of verbal and nonverbal content for repeated human-robot interaction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, Tokyo, Japan, 13-20.
[27] D. Litman, J. D. Moore, M. Dzikovska, and E. Farrow. 2010. Using natural language processing to analyze tutorial dialogue corpora across domains and modalities. *In the Proceedings of the 14th International Conference on Artificial Intelligence in Education*. Brighton, UK.
[28] R. T. Lowe, N. Pow, I. V. Serban, L. Charlin, C. W. Liu, and J. Pineau. 2017. Training end-to-end dialogue systems with the ubuntu dialogue corpus. *Dialogue & Discourse*, 8(1), 31-65.
[29] D. McDuff. 2016. Discovering facial expressions for states of amused, persuaded, informed, sentimental and inspired. *In Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, Tokyo, Japan, 71-75.
[30] D. McDuff, A. Mahmoud, M. Mavadati, M. Amr, J. Turcot and R. E. Kaliouby. 2016. AFFDEX SDK: a cross-platform real-time multi-face expression recognition toolkit. *In Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, San Jose, CA, USA, 3723-3726.
[31] R. Moreno. 2005. Multimedia learning with animated pedagogical agents. *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press, Cambridge, England, 507-523.
[32] S. Namba, S. Makihara, R. S. Kabir, M. Miyatani, and T. Nakao. 2017. Spontaneous facial expressions are different from posed facial expressions: morphological properties and dynamic sequences. *Current Psychology*, 36(3), 593-605.
[33] H. L. O'Brien, and E. G. Toms. 2010. The development and evaluation of a survey to measure user engagement. *Journal of the Association for Information Science and Technology*, 61(1), 50-69.
[34] F. Pecune, A. Cafaro, M. Ochs, and C. Pelachaud. 2016. Evaluating Social Attitudes of a Virtual Tutor. *In Proceedings of the 16th International Conference on Intelligent Virtual Agents*. Springer, Los Angeles, CA, USA, 245-255.
[35] V. Petukhova, T. Mayer, A. Malchanau, and H. Bunt. 2017. Virtual debate coach design: assessing multimodal argumentation performance. *In Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, Glasgow, Scotland, 41-50.
[36] J. Pittermann, A. Pittermann, and W. Minker. 2010. Handling emotions in human-computer dialogues. Springer, Utrecht, Netherland.
[37] V. Ramanarayanan, C. W. Leong, D. Suendermann-Oeft, and K. Evanini. 2017. Crowdsourcing ratings of caller engagement in thin-slice videos of human-machine dialog: benefits and pitfalls. *In Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, Glasgow, Scotland, 281-287.
[38] M. Rehm, E. André, and Y. Nakano. 2009. Some pitfalls for developing enculturated conversational agents. *In International Conference on Human-Computer Interaction*. Springer, Berlin, Heidelberg, 340-348.
[39] L. Ring, D. Utami, and T. Bickmore. 2014. The right agent for the job?. *In Proceedings of the 14th International Conference on Intelligent Virtual Agents*. Springer, Boston, MA, USA, 374-384.
[40] R. Sawyer, A. Smith, J. Rowe, R. Azevedo and J. Lester. 2017. Enhancing Student Models in Game-based Learning with Facial Expression Recognition. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization*. ACM, Bratislava, Slovakia, 192-201.
[41] N. L. Schroeder, O. O. Adesope, and R. B. Gilbert. 2013. How effective are pedagogical agents for learning? A meta-analytic review. *Journal of Educational Computing Research*, 49(1), 1-39.
[42] O. Şerban, M. Barange, S. Zojaji, A. Pauchet, A. Richard, and E. Chanoni. 2017. Interactive narration with a child: impact of prosody and facial expressions. *In Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, Glasgow, Scotland, 23-31.
[43] C. L. Sidner. 2016. Engagement, Emotions, and Relationships: On Building Intelligent Agents. *In Emotions, Technology, Design, and Learning*. 273-294.
[44] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng and C. Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Seattle, Washington, USA, 1631-1642.
[45] K. VanLehn, A. C. Graesser, G. T. Jackson, P. Jordan, A. Olney, and C. P. Rosé. 2007. When are tutorial dialogues more effective than reading?. *Cognitive Science*, 31(1), 3-62.
[46] S. Varges, F. Weng, and H. Pon-Barry. 2009. Interactive question answering and constraint relaxation in spoken dialogue systems. *Natural Language Engineering*, 15(1), 9-30.
[47] M. Wöllmer, M. Kaiser, F. Eyben, B. Schuller, and G. Rigoll. 2013. LSTM-Modeling of continuous emotions in an audiovisual affect recognition framework. *Image and Vision Computing*, 31(2), 153-163.