Modeling Culturally Authentic Style Shifting with Virtual Peers

Justine Cassell Kathleen Geraghty Berto Gonzalez John Borland

Center for Technology and Social Behavior
Northwestern University, Evanston, IL 60208
{justine, k-geraghty, berto, j-borland}@northwestern.edu

Abstract: We report on a new kind of culturally-authentic embodied conversational agent more in line with the ways that culture and ethnicity function in the real world. On the basis of the careful analysis of a corpus of verbal and nonverbal behavior, we found that children shift dialects and ways of using their body depending on social context and task. Based on these results, we implemented a culturally authentic African American virtual peer capable of "code-switching" between African American English and Mainstream American English, and of using nonverbal behavior differently, depending on context. An evaluation of the agent revealed that the virtual peer elicited the same style changes in real children as real children did in one another.

Categories and Subject Descriptors: H.5.1 [**Information Systems**]: Multimedia Information Systems – *Artificial, augmented, and virtual realities.*

General Terms: Language, Human Factors, Design

Keywords: Embodied conversational agent, culture, analysis and modeling of verbal and nonverbal interaction

1 INTRODUCTION

We argue that the implementation of a culturally-authentic embodied conversational agent (ECA) requires a careful analysis of how that cultural identity is displayed through language and nonverbal behavior, and how it is deployed in varying sociocultural contexts. Such an approach, we claim, can result in an agent that is capable of playing real cultural roles, scaffolding important cultural skills, and supporting the construction of identity across cultural contexts. We first motivate the need for a culturally-authentic embodied conversational agent to scaffold the learning of "school English" and school-style science. Then we describe a corpus of data to better understand the use of different dialects, linguistic styles, and science talk in the classroom. We explain how this corpus served as the basis for a probabilistic model for the implementation of an ECA – an African American virtual peer (VP) – capable of collaborating on 3rd grade science tasks. Because TTS and ASR does not exist for African American dialects, our model is implemented in a semi-autonomous Wizard of Oz (WOZ) panel. Finally, we describe the results of a first evaluation, to assess whether children behave similarly with the VP as with their real classmates.

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.

2 BACKGROUND

The Black-White achievement gap is well-known and persistent in the American educational system. According to recent studies, 4th grade Euro American children score, on average, 24 points above the basic skill level in science while African American children score 10 points below basic [27]. By 12th grade African American children's scores have fallen to 26 points below. These scores serve as gatekeepers, barring African American students from the advanced math, science and engineering courses open to their Euro American peers [27]. Effective strategies for diminishing the gap have not been found, and the educational community is hungry for new solutions – especially those that don't require already-scarce human resources. Many have pointed to the role that language plays in the achievement gap [22]. Yet, while traditional science classrooms emphasize a particular style of scientific discourse, not all children come to school having mastered it [19, 21]. These ways of knowing and describing include not just words and grammar, but also the "coordination of appropriate bodily postures, gestures and dispositions" [23] as children learn, for example, how to react to the teacher's and other students' eye gaze and gestures, and what bodily stance is appropriate for classroom interaction compared to play.

As well as being new to particular styles of scientific discourse, not all students come to school speaking the same dialect of English [8]. African Americans may speak a dialect of English known as African American English (AAE), which has its own syntax, morphology and lexicon [12], and may use non-verbal behaviors (such as eye gaze and gesture) in ways that differ from those used by Euro-Americans in similar situations [17]. Different dialects of AAE exist, but many features remain constant. It is important to note that while some African Americans may use AAE forms extensively in all contexts, others may use only some AAE features, and not use others (the deletion of the copula, for example, as in "he ø running" rather than "he is running"). Still other speakers may employ specific AAE features in certain social contexts, but their speech may be indistinguishable from Mainstream American English (MAE) speakers in others.[1] And, of course, not only African Americans use AAE features in their speech, as others may use AAE to signal their identification with aspects of African American identity [10]. The fact that AAE is used by both African American and others should make it clear that AAE, as well as other kinds of communicative nonverbal behavior, can be choices that signal aspects of African American identity [16]. Ethnicity is one aspect of identity, but it is not the only one. In this more socioculturally-influenced approach [30],

[1] The term "Mainstream" American English acknowledges that there exists no single variety of American English that can be called the standard. And yet certain varieties are taken to be the norm, or core of mainstream usage, against which other varieties are measured.

physical appearance is not the only or even the most reliable index of ethnicity. On the contrary, *behavior* may be preferable as an index of identity in ECAs, and "identity" should be considered in the light of the different ways in which we present ourselves.

3 CORPUS COLLECTION

To build a child ECA (a virtual peer, or VP) to support children's acquisition of school-based science talk and MAE in the classroom, we argue that the design process must be informed by research on actual communities of people in particular contexts, and must depend on a model that is derived from the verbal and non verbal behaviors that signal identity in those contexts. After each step of that research, the resulting VP should be brought into classrooms so that we can assess the extent to which it is (a) accepted by children, (b) understood as indexing identity through language use, and (c) is capable of engaging in reciprocal and natural dialogues with children using one variety or other of language available to the children in that classroom. In order to do this, we designed the following data collection procedure.

3.1 Participants

We collected data from 20 African American and 20 Euro American third-grade children (ages 8-10), all of whom live in mid- to low-SES (Socio-Economic Status) neighborhoods (85%-87% low income). Data was collected in 7 schools, Chicago Park District summer programs, and community centers. In order to determine what dialects the sample were capable of using, we initially observed them engaging in a number of different activities, with different conversational partners. The children were asked to describe pictures to AAE and MAE-speaking unfamiliar adults. They were observed in interaction with a familiar MAE-speaking authority figure (the principal), and they were observed in interaction in their school with other children. On this basis, we concluded that one of the African American children spoke MAE in all of the contexts observed, while one of the African American children code-switched from AAE to MAE, depending on interlocutor. The other 22 African American children spoke AAE in all the contexts. All 20 Euro American children spoke MAE throughout. Since our focus in this paper is AAE, this first part of this paper will discuss data from the 20 African American children, from 2 schools and 2 community programs,. Below we will further subdivide the population based on their propensity to code-switch. A later paper will return to all of the children and compare data from the two dialect populations.

3.2 Task

The 20 children in the sample discussed here were matched with a partner from their dialect group and site location, making up 10 dyads and these dyads were asked to complete two tasks: a Bridge and a Classroom task. The resulting corpus represents over 17 hours of data. Because the focus of our efforts is to promote science-talk and MAE use in classroom science activities – without casting aspersions on the child's use of other speech styles or dialects outside of classroom activities – we developed a protocol for a Bridge Task where children first interacted with a partner in a free-play setting, followed by an Classroom Task where the children played the roles of 'student' and 'teacher'.

The bridge task is designed to elicit peer-oriented language in a problem-solving task, whereas the classroom task is designed to elicit formal conventions, science talk, and code-switching from AAE into MAE, if the children do code-switch in formal school situations. Because the children engage in the two tasks with the same peer, any changes in their behavior must be attributed to their sensitivity to the *context*.

In the first, bridge-building, task, the 10 dyads were brought into an empty room, one dyad at a time, where 4 cameras were set up. They were asked to build a bridge out of Lego to span a raging river so as to get three bags of supplies to people (represented by figurines) trapped on the other side. The bags were sufficiently wide and weighted, and the figurines unstable enough so as to make the activity challenging. Children were instructed that the bridge must be long enough to reach over the river, and that they should ensure the bridge was strong enough to support all the bags. After instructions from the researcher, children were left alone to interact, so that no adults would influence their talk. After the children completed the bridge (generally within 15 minutes), the researcher then told the children that they would be bringing their bridge into the classroom and describing their work and their process of problem-solving to the teacher. Children then practiced telling the teacher what they had done, taking turns playing the role of teacher and student with their partners.

3.3 Data Collection

Video and audio were collected from four different cameras, so the entire play area, facial expressions, eye gaze, and body positioning were observable. Data collection took place on location, at the children's schools and after-school programs. Videotapes were digitized and verbal and nonverbal behaviors were transcribed independently. For language, interactions were transcribed orthographically, with the addition of phonetic notation for any non-standard pronunciations or AAE features. All transcribers were familiar with AAE and used Craig et al's [7] inventory of child AAE features. The group of 6 coders included MAE mono-dialectal (African American and Euro American), and MAE-AAE (African American) code-switching speakers. All transcriptions were reviewed by two coders to ensure accuracy.

4 CORPUS ANNOTATION & ANALYSIS

The MAE and AAE dialects and bridge-building, teacher, and student roles represented by the corpus were annotated for a number of linguistic and nonverbal features. Verbal features included: (a) AAE features, including phonological, lexical, and morpho-syntactic, as well as characteristic sing-song intonation (or singing), based on Craig et al. [7]; (b) task-based utterance acts adapted from DAMSL [6], coding of children's science talk [19] and children's role-taking in peer interaction [29]. The coding of non-verbal features included: (c) eye gaze towards partner, task, and away; (d) head movements, including nods, tilts, and the distinctive African American lateral head movement; (f) hand gestures, focusing on iconic and deictic gestures and folded hands; (g) instances of dancing. The co-occurrence of each nonverbal behavior and utterance act was calculated to serve the purposes of our probabilistic model in the system. Our goal in these analyses was to characterize the different contexts – peer oriented problem-solving vs. classroom talk – and different dialects – MAE vs. AAE – in terms of linguistic style, which we take to include both verbal and nonverbal features. This characterization is the model that then serves as the basis for a MAE-AAE code-switching style-shifting VP.

4.1 Linguistic Task Differences

First, to obtain a quantitative characterization of dialect use, we calculated a "Dialect Density Measure" or DDM [7] for each of the children in each task. The DDM calculates the number of AAE features as a rate of total word count. Because AAE shares so many features with MAE, Craig et al. [9] notes that even among the 'heaviest' dialect users, DDM scores are generally around 0.20. In our corpus, DDM scores for third grade African American children ranged from 0.01 to 0.39, with a mean DDM of 0.09. A comparison of bridge to classroom tasks demonstrated that out of the 20 African American children whom we examined, 14 engaged in some degree of code-switching. Since we are building a code-switching VP, it is these children who interest us in the current work. The table below summarizes the mean number of AAE features and words from that group of 14 children, as well as the mean DDMs for those children, and shows that they used 1/3 of the number of AAE features in the classroom task as in the bridge task. Interesting, virtually no difference was found in these children's DDM whether they were playing the role of teacher (DDM of .03) or of student (DDM of .04).

Table 1: Mean Scores in Bridge Building and Classroom Task

		Bridge-building		Classroom	
	(n)	# AAE Features / total words	DDM	# AAE Features / total words	DDM
Code-switching children	14	35/360	0.10	17/508	0.03

These results are striking if one remembers that all of these 14 code-switching children spoke only AAE in every context in which we were able to observe them. They were characterized to us by their teachers as mono-dialectal AAE. Those same children, however, are clearly capable of using multiple language styles – including the use of MAE linguistic features, and a considerable diminishment of AAE features – with a sophisticated notion of context as their guide. It seems that their use of AAE in the classroom – and lack of use of MAE – is a way of indexing identity as is their employing the features of "teacher talk." Encouraging context-appropriate use of MAE in the classroom will involve careful assessment of the reasons –social, developmental, and linguistic - the children have to stick to the use of AAE when interacting with MAE-speaking teachers in classroom contexts. With these results in mind, we set out to examine some of the nonverbal features that distinguished the two tasks for the 14 children who code-switched.

4.2 Nonverbal Task Differences

Results indicated differences in *eye gaze* between the bridge-building and classroom activity. While children engaged in the Bridge Task seldom look at one other, except during moments of laughter, in their roles as either Teacher or Student, children are more likely to look directly at one another, often in combination with a strikingly upright posture, and folded hands. The fourteen code-switching children produced a total of 2352 utterances in total, and each utterance was coded at three positions for gaze direction (Start, Middle, and End of utterance), giving a total of 7056 potential locations for gaze shifts. A chi-square test showed that Role had a highly significant effect on gaze direction at all three positions during utterances: $p. < 0.0001$ for gaze at start, middle, and end. The summed results for beginning, middle, and end are presented in Table 2, below. In terms of probability, children looked at the toy .87 of the time while in the role of Peer, but only .53 of the time while Teacher and .60 as student. They looked at their playmate only 7% of the time while in the role of Peer, but 32% of the time as Teacher and .27 of the time as Student. So, they looked at their peer *three times* as frequently when in the Classroom context.

Table 2: Probabilities of Speaker Gaze Direction by Task*

Gaze target / Role	(n)	Peer	Teacher	Student
Toy	14	0.87	0.53	0.6
Playmate		0.07	0.32	0.27
Elsewhere		0.06	0.15	0.13
Potential gaze shift locations		*4821*	*1002*	*1233*

** Highly significant effect of role on eye gaze target. Descriptive statistics indicate a difference between Peer and Teacher/Student.*

Because children often demonstrate their understanding of physical events and forces through the use of hand gestures [9], we also examined children's hand gestures during the two tasks. As in [9], we found that the children used hand gestures to reinforce and enhance their explanations through demonstrating, framing, and referring. The children used more gestures overall when referring to their bridge, than when they were building the bridge. They were particularly likely to use deictic gestures in the classroom task as they pointed out the aspects of the bridge's structure. Iconic gestures in the classroom context were usually demonstrations**.**

The following examples (Figures 1 and 2), including both speech and nonverbal behavior, illustrate these task-based differences with excerpts from the two tasks of one same dyad.[2]

Spkr	Words *((Actions))*
1	*((Looks at toys))* Lemme see. Lemme see- Yeah **/d**/at's it. It **ø** a fit
2	*((Looks at toys))* **Aight** you make this.
1	*((Looking at toys))* You need, you need another sta/**n**/. Another bigger sta/**n**/ or somethi/**n**/.
1	*((Looking at toys))* We need a stand. You see **/d**/at?
2	*((Looks at toy bucket))* All we need is one more [thing. *((Gets piece, adds to bridge))*
1	[Ok now. There it go **ø**. *((Holds bridge with hands next to 2))*
1	*((Runs hands down sides of bridge))* Dang (.) ((Steps back and claps)) [Yes!

Figure 1: African American Children : Bridge Task

Figure 1 shows significant use of AAE linguistic features, virtually no mutual gaze, and no head nods or other head movements. The classroom task, in Figure 2 below, shows more complex questions, longer sentences, and fewer interruptions; also more MAE features (such as the inclusion of /θ/ rather than /d/), and formal conventions such as *excuse me.* The children in the classroom task are also less physically active, and their gaze focuses on one another rather than the toys. Through both language and nonverbal behavior, the children are making a distinction between the peer-peer and the student-teacher context.

[2] AAE features are indicated in **bold type**

Role	Utterances ((Actions))
T	*((Clears throat))* So S---- *((Looks at S))*
S	*((Looks down, then looks at T, laughs)*
T	*((Looks at S))* Why did you build your side of the bridge the way you built that? *((Looks down))* Why did you do that? *((Looks at S))*
S	*((Leans back in chair))* So. *((Leans forward in chair, replaces hands on table))* So I can get these big o/**l**/ *((old))* bags *((picks up bag))* and put /**um**/ *((them))* on the bridge *((demonstrates putting bag on bridge))* *((Looks at T))* So the bridge can be big and strong for the bags can get over the thing. *((Looks down))* So people can get across and get their [food. *((Looks at T))((Looks at toys))*
T	*((Looks at T))* [Why did you- uh, excuse me. Why did *you ((Looks at toys))* use these two pieces? *((Gets pieces out of bucket))*

Figure 2: African American Children: Classroom Task.
Speaker Roles: S= Student, T=Teacher

5 MODEL

We used the verbal and nonverbal data in the corpus described above as the model for the behaviors of a VP. This VP was designed to engage real children in science discovery and explanation tasks, scaffolding children's exploration of both the use of MAE and the use of science talk in classroom contexts. As we described above, existent approaches to teach MAE to AAE speakers have met with poor success, and the achievement gap in science between African American and Euro-American children remains a serious problem, Meanwhile, although many teachers use only MAE in the classroom, this does not result in increased use of that code in class, as amply demonstrated in our own data where children do not use MAE with adults.

Our hope was that to have a *peer* provide the scaffold would avoid an "oppositional culture" [26] between the child's home and school. We know that children are more likely to pick up languages and dialects from peers than teachers [24], and, the absence of a power relationship such as that between teacher and student (or virtual tutor and child), may increase the child's desire to deploy multiple dialects and linguistic styles. Why *VP*s rather than real peers? In many low SES, ethnically-divided schools there are not always MAE or code-switching speakers with whom to interact. And our prior work has demonstrated that, in fact, VPs are successful at influencing language use – both proto-literacy language [4] and the use of MAE features by AAE speakers [15]. The goal, then, as with our previous work [4], is to introduce the VP into the classroom as a partner in explorations of hybridity and diversity (of language, but also of tools, roles, and social practices) [14]. We do not misunderstand the nature of language use in the classroom – MAE is currently the language in which students are evaluated during, for example standardized testing measures. It therefore seems valid to us to scaffold the child's development of that code – as long as it is seen as one aspect of the child's speech repertoire.

As for the appearance of the VP, unlike previous work on "diversity" in ECAs [*inter alia* 25], we choose both language and nonverbal behavior to signal identity choices, rather than physical appearance. We have shown [15] that ethnicity- and gender-ambiguous VPs may be implemented using an iterative design process until Euro-American and African American children are divided in opinion as to the ethnicity and gender of the agent. Maintaining an ethnicity-ambiguous visual appearance allows us to build an agent that can index identity fluidly through language and nonverbal behavior, rather than being held to a rigid – and perhaps rigidly stereotypical – image of identity defined by the designer's choices of visual appearance.

6 SYSTEM

Essential to a system such as this one is a sophisticated notion of context. The system must know when to use one dialect of English or another, including appropriate nonverbal behaviors. It must know when to use one style of English or another (peer play or teacher-student), along with nonverbal behaviors. And, as will be discussed further below, the system must respond appropriately to the child's physical activities of bridge-building either with collaborative activities, or suggestions, or at least attentive observation. All of these notions of context are currently more-or-less beyond what has been demonstrated autonomously for humanoid agents. So, what follows should be taken as work in progress. However, we believe it important to demonstrate the relationship between context-sensitivity and diversity/culture – the latter being terms which have become quite the mode in the field of ECAs. So, in some places below we will present efforts that have not succeeded. We present them nevertheless because our ultimate goal in the current paper is to demonstrate that VPs that are built on a model derived from a corpus of real children's interactions – though their context sensitivity is due to a wizard behind a curtain – can be as successful as real peers in eliciting context-sensitive behavior in real children.

The linguistic output of the system, in the absence of AAE text-to-speech, was recorded by a native AAE-MAE code-switcher whose voice was pitch-shifted to resemble a child. The "voice talent" recorded all utterances found in our corpus more than a couple of times, and also recorded some modified utterances, such that the system has access to an inventory of utterances with a mean DDM of (0.10) during bridge-building and (0.04) during the classroom task, and the same distribution of particular AAE features (phonological, morphosyntactic, etc.) as found in the child-child corpus. Speech recognition of AAE – or even MAE casual children's speech – is likewise unavailable, and so a human does the speech recognition in our system. But, although we involve a human operator, we want to limit the degree to which his/her adult and subjective judgments influence the VP's behavior. That is, if we want to evaluate whether the VP's peer status plays a positive role in scaffolding language use, then what the VP says should be a reflection of next moves taken from our child-child corpus, as opposed to what the adult operator thinks a child would say. Therefore, as the operator hears the child speak s/he presses a button to assign an utterance type to the real child's incoming speech. The types are taken from those used to annotate the child-child corpus. The inventory of recorded speech is also annotated with this same set of utterance types, and the choice of response utterance is calculated using a Markov model based on what children in the corpus said in a similar utterance act context. That Markov model was generated before run-time on the basis of the annotated corpus, and calculated transitional probabilities[3] for

[3] or conditional probabilities, since the two are equivalent in this case, since the conditional probability of observing state S_1 after state S_2. given that state S_2 has been observed, is equivalent to the transitional probability of moving from state S_1 to state S_2.

each of the utterance categories with which the children's speech was annotated, given a Markov model of the preceding turns in the conversation. Given the utterances of the previous turns as input, and taking into account visual input (taking the place of future Lego sensors) about about the task state, the MM selects the next utterance category based on the distribution found in the corpus. Nonverbal correlates to a given utterance act were automatically chosen from a set of look-up tables for eye gaze, hand gesture, and head movement, based on the co-occurrence probabilities found in the child-child corpus. During an interaction with a child, after the human operator's annotation of incoming speech and action, the WOZ displays a set of highest probability utterance/nonverbal behavior pairs for a response. The WOZ operator then selects an utterance for the VP to produce.

6.1 Predictive/Probabilistic Model

The probabilistic models were calculated by examining the turn boundaries in our corpus, in the contexts of bridge-building, teacher role, and student role. At each turn boundary, the utterance that occurs before the boundary is referred to as the "Given" or G, and the utterance that occurs after the boundary is referred to as the "Target" or T [3]. Considering the utterance categories as states and the turn boundaries as transitions between states, we constructed a first-order Markov model of conversation. Transition probabilities for each pair of utterance categories in the corpus were calculated using equations derived from [3]. In fact, as well as performing a lag-1 sequence analysis, we employed the methods of lag sequential analysis as described in [3] to calculate the equivalent of skip-grams [13]. That is, we predicted Alex's utterance on the basis of events that are not necessarily temporally adjacent. These skip-bigrams should have a comparable rate of incidence to standard bigrams in the data, therefore allowing longer histories to be considered without exponential growth of training data. In the long run, however, this kind of non-sequential transitional probability calculation seems unsatisfactory, and we are currently experimenting with third-order Markov chains with a way of backing off to a lower-order model when data are too sparse to justify a third-order model.

6.2 SmartBody & PandaBMLR

Graphics rendering engines come and go, as do TTS, and ASR. In order to be able to substitute state-of-the-art modules as they are developed, our current architecture seeks the highest level of abstraction possible for modeling human behavior within the decision module. To this end, as much computation as possible is offloaded from the graphics engine, with motion control and realization separated. Motion control is offloaded to SmartBody [28], specifically designed to translate high-level descriptions of behavior into synchronized joint rotations. The high-level descriptions of behavior are sent to SmartBody in Behavior Markup Language. The visualization of the virtual agent is handled by PandaBMLR [2], which takes the output from SmartBody and renders it in a lifelike environment with Panda3D [11], an open-source rendering package. This allows, for example, high-level descriptions of gestures to be translated into the correct input for the given rendering engine.

6.3 Shared Reality

For the system at hand, in a continuation of the "shared reality paradigm" that defined our very first VP [5], we are implementing a touch surface table that detects Lego blocks using the WiiMote method [20]. The child stands at one edge of a table, while the VP is projected onto a large screen facing the child, appearing to stand at the other end of the same table. During an interaction, the child and VP co-build a bridge – so that the two halves meet on an island in the middle. As can be seen in the screenshot, the VP (referred to by the gender-neutral name of Alex) has access to a bin of virtual Lego bricks. The child has access to a bin of identical physical Lego bricks, as well as figurines and a weight to represent the bag of food and supplies. The VP chooses and grasps objects that are similar to the objects the real child is manipulating. To achieve this goal, using a proposed new BML tag , <grasp>, when Alex reaches into a bucket, a block attaches itself to a joint on Alex's hand. With the block in hand, Alex can then perform building actions, show the block to the child, and so forth. This same technique is used during the teacher-student phase of the exercise where Alex places figurines or weighted bags on the bridge to assess the strength of the bridge. When the block in Alex's hand reaches the point where it needs to be placed, the block is rendered invisible and a new block appears on the table. Alex's wrist is bent so the palm points inward blocking the child from seeing the 'magic' of how blocks are placed. The blocks are also tagged so as to act as gaze targets so that Alex can watch the child build. These achievements make it appear as though Alex is building along with the child in real time, creating a collaborative, shared-reality environment. Note that two different kinds of gestural multimodality must be scheduled with respect to one another: the *actions* of bridge-building, and the *co-verbal gestures* that children use in demonstrating what the other child should do, or describing what they have built. The two systems run as separate parallel processes, but co-verbal gestures have scheduling priority such that the VP may pause while placing a brick, perform an unrelated co-verbal gesture with the brick in its hand, and then return to building the bridge.

7 EVALUATION

End-to-end dialogue systems are notoriously difficult to evaluate – all the more so if they include multiple modalities. We therefore chose to begin by simply assessing whether the system evoked as much or more use of MAE and science talk as did real peers. If they do, this lays the ground work for employing them in classrooms where there exist no peers to serve as code-switching models nor who employ classroom-ratified scientific discourse. This choice was based on the assumption that the semi-autonomous WOZ panel played a role in rendering the vp as believable. Clearly, however, the probabilistic model method of building a WOZ, and the efficacy of this particular multimodal system, in encouraging code-switching, and formal science talk will need to be subsequently evaluated more rigorously through comparisons of different versions of the system.

In this initial evaluation six AAE-speaking 3rd grade students from a 100% African American school participated in interactions with the VP, giving us 6 dyads to observe. To establish a base rate for the children's DDM, children first engaged in a pre-task picture description with the researcher, a Euro American, MAE-speaking adult. This task has been found to be a robust and reliable

measure of DDM, regardless of the ethnicity of the adult [8]. Then, the experimenter introduced the bridge-building task. The child and Alex were asked to build a collaboratively build a bridge towards an island in the middle. Alex's non-verbal behaviors included gaze direction, head movements, and some hand gestures, in addition to bridge-building and strength-testing actions. As in the child-child data collection, no adults were present during the task, and after children built the bridge, the researcher re-entered the room and explained the second part of the task to the child and Alex, designating the real child first in the role of the 'Student.' The child then had five minutes to play the role of student, and five minutes to play the role of teacher. Alex's speech in the bridge task demonstrated a DDM of .10, and a DDM of .01 in the classroom task.

7.1 Results

We were gratified by the speed and accuracy with which the human operator actually could tag the correct utterance act for the child's incoming speech, and choose utterances for Alex to speak. What Alex actually spoke in response, however, was more of a problem, for two reasons. On the one hand, our application of Bakeman and Gottman's [3] lag sequence analysis ultimately resulted in the system predicting the Target sometimes on the basis of the previous Given (lag 1), sometimes on the utterance two turn boundaries back from the Target, and sometimes on the utterance three turns back. Clearly there is some kind of relationship between each of these Givens and the Target, but it is not linguistically motivated in the way a response-pair should be.

On the other hand, due less to the Markov model and the semi-autonomous mode than the generic nature of our original utterance type coding, we found that Alex's utterances were often off-target. While the Lego sensors (inputted by the human operator at each task stage) filtered responses so only context-appropriate utterances (no bridge, partially-built bridge, testing weight phase), sometimes none of the questions presented to the operator had anything to do with the current sub-goal – a granularity problem due to our coding scheme. For this reason, we are now re-coding our entire child-child corpus with a new annotation scheme, designed specifically for this micro-world, and intended to give us a sense of what specific bridge-building goals (for the bridge task), and what kinds of justification, motivation, and explanation the child is currently engaging in (for the classroom task). Our worry, of course, is that this greater level of detail will yield a sparse distribution, and will not provide sufficient power to calculate z-scores for the Markov model. For this reason, each child's utterances is calculated with a supra-goal, sub-goal, utterance type, and topic, so that we can back off a level if we don't have enough instances of the finest granularity of coding. To avoid that eventuality we are now also collecting and integrating twice as much data into our corpus.

However, it does seem as if the language and nonverbal behavior of the VP did contribute to its believability, as all of the children who participate were highly engaged, singing with Alex, talking back, and in other ways demonstrating their engagement. More to the point, results concerning Alex's ability to encourage code-switching and formal science talk were very positive.

A one-way, repeated measures ANOVA[4] showed a significant main effect of Task on Mean Length of Utterance (MLU, a common index of complexity in children's speech), $F(2,15) = 3.16$, $p<0.001$, with significantly longer MLU in the Classroom task than in the Bridge Task, indicating increased complexity of utterances (see Table 3, below). The role of the child also had a significant effect on MLU, with 'teachers' speaking in a more complex way than 'students', On the other hand, unlike the child-child corpus, there was only a trend towards significance for the effect of Task on overall DDM, $F(2,15) = 2.762$, $p=0.08$). Below, the table shows the mean scores, subdivided into each Role. The children's mean DDM is highest in the Picture Description and lowest in their role as Teacher. Thus, as we found when observing our initial population, paradoxically, the children are most likely to speak AAE with MAE-speaking adults. And, to our gratification, they are least likely to speak AAE when playing the role of an MAE-speaking teacher.

Table 3: Mean scores for Task and Role in Child-VP Sessions. ^
Significant main effect of Task on MLU, $p < .05$
^ Picture description task with MAE-speaking, Euro American Adult

	Picture ^	**Building**	**Classroom**	
Measures	***Child^***	*Peer*	*Teacher*	*Student*
MLU	6.12*	3	7.98*	5.73*
DDM	0.171	0.164	0.042	0.084

To further explore the trend for the effect of Task and Role on DDM, we divided the children's AAE features into morpho-syntactic (MorDDM) and phonological (PhonDDM) and compared the bridge building and classroom tasks. With this more detailed level of analysis, Task showed a significant main effect on MorDDM, $F(2,21)=4.872$, $p<.05$, but there was no main effect of Task on PhonDDM, $F(2,21)=2.251$, $p=0.13$. Role also showed a significant effect on MorDDM, $F(3,20)$, $p < .05$.

Table 4: DDM Measures in Child-VP Sessions
^ Picture description w/ same MAE-speaking, Euro-American Adult
** = Significant at $p > .05$, †= Marginally significant*

	Picture	**Building**	**Classroom**	
DDM Measure	*Child^*	*Peer*	*Student*	*Teacher*
DDM	0.171 †	0.164	0.084	0.042 †
MorDDM	0.032*	0.0075*	0.015*	0.004*
PhonDDM	0.139	0.160	0.069	0.038

Post-hoc t-tests showed that the difference between mean DDM for Picture Description and Teacher was marginally significant, $t(23) = 2.086$, $p=0.576$. For MorDDM, the differences between Teacher and Child was significant, $t(23)=2.086$, $p < 0.01$and Student and Child was significant, $t(23) = 2.086$, $p < 0.05$. Comparison of the PhonDDM showed that the difference between Peer and Teacher was marginally significant, $t(23) = 2.086$, $p = 0.071$. Most striking is that children are using virtually no AAE morphosyntax and least AAE phonology in the teacher role. With a larger *n*, future work will entail more detailed linguistic analysis of specific features.

7.2 Sample VP-Child Transcripts

Examples illustrate these phenomena. Figure 3 (below), shows an interaction between the VP and AAE speaking child during bridge-building. Both child and VP use phonological,

[4] Repeated measures ANOVA is motivated by the fact that the same VP interacts with each child, and therefore since we are only looking at the real child, this is not truly dyadic data.

morphological, and syntactic, AAE features. As evidenced in the MLU scores, and similarly to the child-child corpus, speaking turns in the Child-VP interactions are relatively short.

1	Alex	This block, **it ø** just right. *((Shows block, puts block on bridge))*
2	Child	Hmm *((Looks at bridge))*
3	Alex	Whaddyou think we should do? *((Looks at bridge, then looks at child))*
4	Child	I think that we **/shou/** put the long ones behind here- behind the bridge (.) and then stack some over here to make some stairs ((Starts attaching more pieces))
5	Alex	There it **go ø** *((Looks at child's side of the bridge))*
6	Child	Now Alex we're gonna /s/- We **ø** gonna try (.) to see if (.) it hold **ø** up all the food *((touching pieces along the top of bridge))*

Figure 3: Child-VP Bridge Building Interaction

In contrast, during the Classroom Task, (Figure 4, below) longer turns were characteristic, The following example is from the same child, with the VP acting as Teacher. The child uses longer and more complex sentences, and uses only one AAE feature (zero past tense) during this long segment, with a DDM almost half what it was during her bridge-building task). Notice that although she expresses her understanding of causal relationships (e.g. the bridge falling), the answers have little to do with the science of the task (e.g. weight, forces, testing her design),. A VP that can model evidence-based scientific justification may be able to bring this child one step closer to classroom-ratified science talk.

Alex	*What was the problem you had to solve today?*
Child	Okay
Child	The problem I **have** to solve (.) Alex was- the problem I **have** to solve Miss-Miss Alex was to get the people over there to the bridge so they could reach the food
Child	And so the fishes- so the fish would not get them
Child	And they /ha/- because they didn't- because they didn't have any food at all
Child	So I made this bridge to carry over and back so they can walk over it
Child	And I accomp- accomp- accomplished our (.) mission
Child	I made the bridge

Figure 4: Child-VP during Classroom Task: Child playing student

8 DISCUSSION AND FUTURE WORK

The results of this evaluation are highly encouraging for the use of VPs as a way to both model and elicit code-switching from AAE-speaking children, and for the use of VPs as a model for children's science talk. The increase in the child's MLU in the Classroom Task, reduction in DDM, as well as qualitative analysis of the interactions, show that students are able to adjust their speech to the appropriate (academic) social context. What is particularly interesting about the reduction in DDM is that none of the participating children had been identified as code-switchers, and yet *all* of them showed the ability to reduce their DDM based on the social context. As with the child-child corpus, our observations indicated that these children were mono-dialectal AAE speakers –not surprising given that the highest percentage of AAE features that they used were when speaking with an adult!

The children's language showed that they were attentive to the structure and style of Alex's talk, albeit differently in the two tasks. For example, during bridge building children sang along if Alex sang. During the classroom task, when children played the role of teacher, they often repeated or re-formulated questions that Alex had asked when Alex was the teacher. Children also reformulated Alex's teacher questions when they took the role of student. Thus, the child appears to view Alex as competent in the genre of classroom science talk. Though qualified by the number of participants, the results are encouraging, and suggest that the VP is capable of modeling MAE use, classroom science talk, and even affecting children's code-switching behavior.

Our future work includes our current additional child-child data collection, and use of the new annotation scheme we are currently applying. Additional analysis of children's non-verbal behaviors, specifically in regards to hand movements and other full-body movements will inform our continued development of the VP, such as the co-occurrence of specific AAE features with specific non-verbal behaviors. The annotation scheme will also be applied to our transcribed MAE dataset so that comparative analyses are possible. Ultimately we hope to have significant corpora from mono-dialectal MAE and AAE speakers, and code-switching speakers, and we plan to analyze the effects of interaction with the VP for each of these populations, with a much larger *n* than was possible for this initial study. In addition, we are planning an evaluation of the probabilistic model by, first, implementing a Katz Backoff Model with Good-[18] estimation as described in to address the problem of sparse datasets, and then comparing that VP to one that is controlled by an adult without the support of the child-child corpus.

Technically, we are currently implementing the Lego sensors to give information about the state of each task. We also intend to implement eye gaze detection, which we have used successfully in the past to improve the interaction between the real and virtual participant. In conjunction with our continued belief in one day finding ASR and TTS for dialectal English, we believe that we will ultimately be able to place Alex in the classroom as a real partner in learning.

9 CONCLUSIONS

Our results suggest that children of this age group are subtly sensitive to the social contexts of the classroom, and able to adjust their linguistic and nonverbal behavior accordingly. This sensitivity, however, appears to be linked as much to the children's exploration of identity, as to their desire to succeed at a school task. In addition, while our attempt to design a semi-autonomous WOZ interface on the basis of these results was only partially successful, it is still the case that VPs designed on the basis of careful study of those contexts, and of the linguistic and nonverbal phenomena that appear in them, seem to be able to induce code-switching, as well as model appropriate science talk – the latter somewhat better than the real peers do. This is a striking result given the lack of success that teachers have in inducing the use of MAE in the classroom, and it suggests that this is one place where VPs might succeed better than intelligent tutors – which most often resemble teachers and not peers. In sum, these results suggest a role for authentic, enculturated multimodal virtual peers.

10 ACKNOWLEDGEMENTS

This work is generously supported by the National Science Foundation Award #073664. The authors wish to thank the members of the ArticuLab and the CollaboLab at Northwestern University, the parents, teachers, staff and children involved in the project, and our research assistants: Shiana Crosby, Amy Cleveland, Angelica Cleveland, Breyana Drew, Courtney Gilliam, Evelyn Parks, and Katy Witmer.

9 REFERENCES

[1] Allen, B. A. and Boykin, A. W., (1991) "The Influence of Contextual Factors on Afro-American and Euro-American Children's Performance: Effects of Movement Opportunity and Music.," *International Journal of Psychology,* vol. 26(3), 373-388.

[2] Arnason, B. T. and Thorsteinsson, A. (2008) *BML Realizer: An Open Source BML Animation Toolkit*. Unpublished Bachelor of Science dissertation, School of Computer Science, Reykjavik University.

[3] Bakeman, R. and Gottman, J. M., (1997) *Observing Interaction: An Introduction to Sequential Analysis*, Second ed.: Cambridge University Press.

[4] Cassell, J., (2004) "Towards a Model of Technology and Literacy Development: Story Listening Systems," *Journal of Applied Developmental Psychology,* vol. 25(1), 75-105.

[5] Cassell, J., et al., (2000) "Shared Reality: Physical Collaboration with a Virtual Peer," in *Proceedings of CHI 2000*, The Hague, The Netherlands, 259-260.

[6] Core, M. and Allen, J., (1997) "Coding Dialogue with the DAMSL Annotation Scheme," in *Proceedings of AAAI Fall Symposium on Communicative Action in Humans and Machines*, Boston, MA.

[7] Craig, H. K., et al., (2003) "Phonological features of child African American English," *Journal of Speech, Language, and Hearing Research,* vol. 46(623-635.

[8] Craig, H. K. and Washington, J. A., (2005) *Malik goes to school: Examining the language skills of African American students from preschool-5th grade*. New York: Lawrence Erlbaum.

[9] Crowder, E. M., (1993) "Telling What They Know: The Role of Gesture and Language in Children's Science Explanations," *Pragmatics and Cognition,* vol. 1(2), 341-376.

[10] Cutler, C., (2003) ""Keepin' It Real": White Hip-Hoppers' Discourses of Language, Race, and Authenticity," *Journal of Linguistic Anthropology,* vol. 13(2), 211-233.

[11] Goslin, M. and Mine, M. R., (2004) "The Panda3D Graphics Engine," *Computer,* vol. 37(10), 112-114.

[12] Green, L. J., (2002) *African American English: a linguistic introduction*. Cambridge, UK: University of Cambridge Press.

[13] Guthrie, D., et al., (2006) "A Closer Look at Skip-gram Modelling," in *Proceedings of Fifth international Conference on Language Resources and Evaluation (LREC)*, Genoa, Italy, 1222-1225.

[14] Gutiérrez, K. D., et al., (1999) "Rethinking Diversity: Hybridity and Hybrid Language Practices in the Third Space," *Mind, Culture, and Activity,* vol. 6(4), 286-303.

[15] Iacobelli, F. and Cassell, J., (2007) "Ethnic Identity and Engagement in Embodied Conversational Agents," in *Proceedings of 7th International Conference on Intelligent Virtual Agents*, Paris, France, 57-63.

[16] Irvine, J. T., (1985) "Status and Style in Language," *Annual Review of Anthropology,* vol. 14(577-581.

[17] Johnson, K. R., (1976) "Black kinesics: Some non-verbal communication patterns in the Black culture," in *Intercultural communication: A reader*, L. A. Samovar and R. E. Porter, Eds. Belmont, CA: Wadsworth, 259-268.

[18] Jurafsky, D. and Martin, J. H., (2006) *Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing*, 2nd ed.

[19] Kurth, L. A., et al., (2002) "Student use of narrative and paradigmatic forms of talk in elementary science conversations," *Journal of research in science teaching,* vol. 39(9), 26.

[20] Lee, J. C., (2008) "Wiimote Tracking," Pittsburgh, PA

[21] Lee, O., (1999) "Equity Implications Based on the Conceptions of Science Achievement in Major Reform Documents," *Review of educational research,* vol. 69(1), 83.

[22] Lubienski, S. T., (2001) "Are the NCTM "Standards" Reaching All Students? An Examination of Race, Class, and Instructional Practices," in *Proceedings of Annual Meeting of the American Educational Research Association*, Seattle WA, 22.

[23] Luke, A., (1992) "The body literate: Discourse and inscription in early literacy training," *Linguistics and Education,* vol. 4(1), 107-109.

[24] Mashburn, A. J., et al., (2009) "Peer Effects on Children's Language Achievement During Pre-Kindergarten," *Child Development,* vol. 80(3), 686-702.

[25] Nass, C., et al., (2000) "Truth is Beauty: Researching Embodied Conversational Agents," in *Embodied Conversational Agents*, J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, Eds. Cambridge, MA: MIT Press, 374-402.

[26] Ogbu, J. U., (2003) *Black American students in an affluent suburb: a study of academic disengagement*. Mahwah N J: L. Erlbaum.

[27] Perie, M., et al., (2005) "The Nation's Report Card: Reading 2005," National Center for Education Statistics (NCES), Ed.: US Department of Education.

[28] Thiebaux, M., et al., (2008) "SmartBody: behavior realization for embodied conversational agents," in *Proceedings of 7th international conference on Autonomous agents and Multiagent Systems*, Estoril, Portugal, 151-158.

[29] Wang, A. and Cassell, J., (2003) "Co-authoring, Corroborating, Criticizing: Collaborative Storytelling for Literacy Learning," in *Proceedings of Vienna Workshop: Educational Agents - More than Virtual Tutors*, Vienna, Austria.

[30] Wertsch, J. V., et al., (1995) *Sociocultural studies of mind*. Cambridge ; New York: Cambridge University Press.