Dialogue & Discourse 10(1) (2019) 20–33 doi: 10.5087/dad.2019.102

Using Gestures to Resolve Lexical Ambiguity in Storytelling with Humanoid Robots

Catelyn Scholl CJSCHOLL@UWM.EDU
Computer Science
University of Wisconsin-Milwaukee

Susan McRoy MCROY@UWM.EDU
Computer Science
University of Wisconsin-Milwaukee

Editor: Manfred Stede

Submitted 02/2018; Accepted 01/2019; Published online 02/2019

Abstract

Gestures that co-occur with speech are a fundamental component of communication. Prior research with children suggests that gestures may help them to resolve certain forms of lexical ambiguity, and might also help them learn homophones. To test this idea in the context of human-robot interaction, the effects of iconic and deictic gestures on the understanding of homophones was assessed in an experiment where a humanoid robot told a short story containing pairs of homophones to small groups of young participants, accompanied by either expressive gestures or no gestures. Both groups of subjects completed a pretest and post-test to measure their ability to discriminate between pairs of homophones and we calculated aggregated precision. The results show that the use of iconic and deictic gestures aids in general understanding of homophones, providing additional evidence for the importance of gesture to the development of children's language and communication skills.

Keywords: lexical ambiguity, speech prosody, homophones, iconic gestures, deictic gestures

1. Introduction

We have been using humanoid robots to teach children about computers, showing them how programs can control how robots move, speak, or play music. We have noticed that children across many age groups were more attentive to the robot sessions than to other demonstrations of computers. We have also noted that these children were not especially interested in complex spoken interaction with the robot on its own and have wondered about the potential impact on attention or learning if we were to add coordinated gestures to the robot's spoken behaviors. It is well known in human interaction that using gesture is not just a matter of naturalness; body movement or gestures are an important form of communication, especially for young children or new learners of a second language (Kidd and Holler, 2009; Ray, 2015). At some point, teachers might want to incorporate robots into their classrooms as a way of enhancing student's language skills. Thus, it would be important to assess the potential benefit of adding gestures to robot-human communication with children.

©2019 Catelyn Scholl and Susan McRoy
This is an open-access article distributed under the terms of a Creative Commons Attribution License
(http://creativecommons.org/licenses/by/3.0/).

Our study attempts to measure the potential benefit of adding gesture to robot speech to help discriminate among homophones, where the target population is young children who are not yet fully fluent in their own language. The assessment involves storytelling, which would be a normal activity for this age group and also something that one might want to incorporate in classroom activities. Homophones occur when two words are pronounced alike but have different meanings, derivations or spellings. An example of a homophone would be *blue*, the color, and *blew*, the past form of the verb *blow*. Given this task, we then consider what types of gestures might be helpful and how these gestures can be timed appropriately with speech.

1.1 Background

Gestures come in many types. Some gestures are used instead of speech and some are meant to enrich it. For example, Ekman and Friesen (1969) define five types of gestures for communication of content and emotion, as well as for non-communicative functions. These types include emblems (for directly replacing some words), illustrators (for shaping what is being said), affect displays (for showing emotion), regulators (for controlling the flow of conversation) and adapters (for helping to relieve the speaker's own tension). If one focuses on co-speech gestures, the gestures that accompany and enrich speech rather than replace it, these are most often divided into four subtypes, following McNeill (1992). These four subtypes are: iconic (for representing object attributes), deictic (for relating one entity to another by pointing), metaphoric (for relating an abstract idea to a concrete one by creating a picture) and beat gestures (for keeping the rhythm of the speech, without expressing content.) Furthermore, according to McNeill, these sorts of gestures and speech involve the same mental process and thus are naturally temporally and semantically related. Indeed among children less than 18 months old, vocabulary acquisition has been found to be better when words were accompanied by appropriate iconic gestures than when used without gesture or with an uncoordinated gesture (see Namy et al., 2000; Zammit and Schafer, 2011).

Gestures can help convey meaning by helping listeners to resolve lexical ambiguity or to better understand and acquire unfamiliar vocabulary. Gestures may also enhance attention by making the communication livelier or more memorable. The use of gesture is one of the first forms of intentional co-ordination of behavior that people learn. Gesture appears to precede early verbal language development and it has been surmised that the use of pointing gestures may act as a precursor to initial acquisition of single words, so as children point at things, people tell them what they are called. Children also combine multiple gestures representing different concepts before producing two-word utterances, (Kidd and Holler, 2009).

Past work with young children and the use of gesture during storytelling has focused on the use of deictic and iconic gestures, as they convey meaning but do not require a complex mapping between abstract and concrete concepts. For example, Kidd and Holler (2009) considered children's own use of gesture while retelling a story containing homonyms, which is where the same word can have two very different meanings, for example *bat* as an animal versus as an object. Their study involved three to five -year-old children. For the study, the researchers created short story books containing age-appropriate homonyms such as *bat*, *mouse*, and *glass*. The stories were read to the children and then the children were asked to retell the story while the researchers considered how they resolved the homonym ambiguity. The results revealed that the youngest children, 3-year-olds, primarily used nouns and deictic gestures as their means of disambiguation. The 4-year-olds added more iconic gestures. The 5-year-olds mainly relied on speech alone.

Another study, involving children who were telling stories, investigated the use of gesture in resolving lexical ambiguity among English-as-a-Second-Language (ESL) learners (Ray, 2015), The results suggest that these learners frequently used gestures when telling stories, and younger age groups used gesture more frequently than older groups. The main difference between children who are second language learners versus first language speakers seems to be that second language learners, especially those who appeared to have the most difficulty with speaking the second language, used iconic gestures more frequently than deictic gestures.

No work has been done to measure the use or impact of gestures on lexical disambiguation for robot-human communication with children. One study, by Kory Westlund et al. (2017), assessed the importance of speech prosody on children's acquisition of new vocabulary and their ability to recall and retell a story. The study compared the impact of "flat" (non-prosodic) versus "expressive" (with prosody and gestures) storytelling. Their findings suggest that children were more effective in using newly acquired vocabulary after listening to the expressive robot. Furthermore, during a delayed recall test, children who listened to the expressive story had less difficulty retelling the story than those who had heard the flat storytelling. There is another gesture-related study with robots (see van Dijk et al., 2013) that considers the impact of robot-generated iconic gestures during interaction with older adults. That study found that the gestures helped in the retention of information. Gestures may help all listeners (old and young) in retaining information about a story including the semantic context necessary to understand lexical ambiguities such as homophones.

1.2 Overview of the Present Study

The goal of our work was to assess the hypothesis that robot-to-human speech accompanied by gesture would improve children's ability to disambiguate homophones more than without gestures. We conducted an experiment using humanoid robots to tell stories to preschool and early elementary age children and assessed the children's ability to discriminate homophones both before and after experiencing the story. Our experiments included only iconic and deictic gestures because they carry the most concrete semantic information and thus are likely to be the most understandable to young children.

We selected an existing story from the children's literature appropriate for the ages of the subjects recruited for the study. To make the storytelling more natural, in both conditions we implemented the tempo aspects from an existing rule-based model of speech prosody for storytelling when converting the text to speech (Theune et al., 2006). This approach ensures that the storytelling is not too fast for the anticipated skill level of listeners and that the boundaries of phrases and sentences are clear. Our implementation does not put any stress on any of the words, including the homophones, that might draw attention to them.

2. Methods

We used a humanoid robot, programmed to read the published children's story *Dear Deer: A Book of Homophones*, (Barretta, 2010). Table 1 includes some sample sentences from this work. Where possible, we added simultaneous movements by the humanoid robot corresponding to different related gestures. For assessment, we created a pretest and a post-test using a subset of the vocabulary from the story and wrote our own simple sentences. After approval by our Institutional Review Board, we recruited subjects from a university-run daycare center and conducted our experiment on site. More detail is provided below.

The MOOSE loves MOUSSE.
He ATE EIGHT bowls.
Have YOU seen the EWE?
She's been in a DAZE for DAYS.
The TOAD was TOWED to the top of the seesaw, so he could SEE the SEA.
The WHALE was ALLOWED to WAIL ALOUD.
The BEAR had to PAUSE to BARE his big PAWS.
HEY, the elephant THREW a pail THROUGH the big bale of HAY!
Have you READ about the RED fox who BLEW BLUE bubbles?

Table 1: Sample sentences from *Dear Deer*

2.1 The Robot Framework

Two separate presentations were created to tell the story. The first presentation was for the control group, which contained only speech for the homophone pairs. The second presentation, for the experimental group, contained speech and gestures that corresponded with some of the homophones.

We used a NAO robot and its accompanying software, called Choreograph, to implement the story teller. Choreograph software allows one to create timed sequences of actions by the robot, including simultaneous physical motions and speech. Some poses, such as sitting, are provided as part of the library. Choreograph software also supports something called animation mode, which allows one to define new actions using the real robot like a puppet. In this mode, one can manually move any part or set of parts and save the joint settings as an object, called a "box". Then, one can create a sequence of boxes called a "timeline" that combines predefined behaviors or poses (such as "say" or "sit") and ones created using animation mode. One can also control the duration of each behavior, so that a particular movement or spoken phrase can be synchronized (e.g. by having them start at the same time and have approximately the same duration). Once set, the robot will perform the story the same way each time.

Each presentation (speech and gesture) was created using a combination of functions from the Choreograph library and some additional python scripts. Specifically we used the existing functions to control text-to-speech, but added some python scripts to calculate the correct timing for speech and pauses, based on Theune's model. A few physical motions (like sitting, and opening or closing a hand) were used as predefined, but most required the use of animation mode to approximate the gesture. Then all the actions were manually placed in a sequence. Adjustments to the timing of the gestures in a timeline were made by watching the robot tell the story and manually adjusting the timing to provide the most natural observer experience. The final parameterized timelines (with and without gestures) were saved as separate objects, that when initiated, cause the NAO robot to present the complete story without additional interaction. In both conditions, the robot begins in its initial position (a crouch), and moves to a sitting position to tell the story.

2.2 Speech Prosody

To implement more natural sounding phrasing when converting text from the short story to speech, we used part of a rule-based framework for storytelling prosody (Theune et al., 2006). This work provides rules for a global storytelling speaking style, which specify a consistent and natural set of speech variations, where the goal is to better emulate natural storytelling versus read speech. We used only the rules for controlling the timing of phrases and pauses, both to seem more natural and also to help clarify the syntax.

We implemented these rules directly in the text-to-speech API for NAO robots, adjusting the duration of syllables. All other aspects of speech, such as pitch or stress on vowels or syllables, were controlled by the standard API. We did not implement Theune's other rules for pitch or intensity as these aspects seemed adequate for our purposes. We did not make any specific adjustments for homophones, as in some sentences of the story nearly every word is a homophone (e.g. *He ate eight bowls.* and *The whale was allowed to wail aloud.*).

The timing rules, as we implemented them, are summarized below:

1. A rate of about 3.6 syllables per second was used for the tempo. This was found by counting the syllables in a few trial sentences and averaging the time it took to read the sentences.

2. The duration of accented vowels was 1.5 times their average duration. This was done using NAO's text-to-speech API.

3. For pauses, after each phrase, we added a break of a length of 0.4s and between sentences we added a break of length of 1.3s. These pauses were also created using NAO's text-to-speech API.

2.3 Gestures to Accompany the Homophones

The children's story *Dear Deer: A Book of Homophones* (Barretta, 2010), is a story created with the intention of introducing homophones, by incorporating pairs of homophones with pictures and sentence content to help discriminate the respective meanings. Telling the story typically takes an adult three to five minutes (see Jean, 2015). The story is 241 words long and has a reading Lexile Level of 530L and an ATOS Reading Level of 2.1.i which means that it is considered appropriate for a seven year old to read independently (TeachingBooks.net, 2018).

The story contains the 32 homophone pairs listed in Table 2, shown with their primary part of speech. Each sentence of the story contains from one to three homophone pairs (most contain two). Most pairs occur only once except for "Dear-Deer" and "Aunt-Ant", which are each repeated once.

Manual review of the pairs suggested that for 11 of the pairs at least one homophone sense used in the story had a possible iconic or deictic gesture and one pair had two, one for each sense. Four of the homophone senses involved actions for which there is an associated body part (Hear-ear, See-eye, Ate-mouth, Wail-eyes). Three were a body part (Hair, Feet, Paws). Three seemed naturally deictic (Here, You, Him). One other was an animal with a distinctive feature (Moose-animal with big antlers) and the other an adjective associated with a body part (Hoarse-throat).

The specific gestures were chosen to be ones that the authors felt a human reader might use, based on the lexical meaning of the homophone and the typical way that such meanings are depicted. Past research suggests that deictic and iconic gestures are used by even very young children and might include gestures like pointing towards a foot to indicate the body part, feet (see Iverson et al., 1994). The gestures chosen were also ones that could be performed given the capabilities of the humanoid robot. The flexibility and motor control of the robot limit it to moving its arms or hands in different directions (left, right, up, down), at various angles, and can be positioned manually to reach towards one of its parts (head, feet, eyes, mouth, ears, neck, torso) and either holding the appendages still (hold, point), rotating them slightly to simulate rubbing (rotate) or letting them

allowed (v)	aloud (adv)	hey (adv)	hay (n)
ate (v)	eight (n)	him (n)	hymn (n)
aunt (n)	ant (n)	horse (n)	hoarse (adj)
bear (n)	bare (adj)	kneaded (v)	needed (v)
bee (n)	be (v)	know (v)	No (adv)
blew (n)	blue (adj)	mood (n)	mooed (v)
choose (v)	chews (v)	moose (n)	mousse (n)
daze (n)	days (n)	news (n)	gnus (n)
dear (adj)	deer (n)	paws (n)	pause (n,v)
doe (n)	dough (n)	read (v)	red (adj)
feat (n)	feet (n)	sea (n)	see (v)
flew (v)	flu (n)	tale (n)	tail (n)
flea (n)	flee (v)	threw (v)	through (prep)
hair (n)	hare (n)	toad (n)	towed (v)
hear (v)	here (pron)	whale (n)	wail (n,v)
herd (n)	heard (v)	you (pron)	ewe (n)

Table 2: Pairs of homophones, with parts of speech for each, where *n* is *noun*, *v* is *verb*, *pron* is *pronoun*, *prep* is *preposition*, *adj* is *adjective*, and *adv* is *adverb*.

drop to end the gesture (move). If there was no plausible, achievable gesture then the arms of the robot would remain down and still during that part of the sentence.

Overall, 11 pairs were chosen and 12 corresponding iconic or deictic gestures were selected. Table 3 provides a dictionary of gestures for the homophones that were identified as having a corresponding gesture and a brief description of the gesture that was used. The robot was programmed to implement the gesture associated with the word so that it would occur at approximately the same time as the corresponding word was spoken.

The timing of the gestures was verified by videorecording the robot telling the story with the gestures and asking adult observers to view the video (on Youtube) and judge the naturalness of the timing of the gestures. Each judge was asked to rate the timing according to the following rubric:

The timing of the gestures was poor they were distracting to watch. The story would be better with no gestures at all.

The timing of the gestures was okay a few times they were noticeably too early or too late.

The timing of the gestures was good they seemed natural, but at least once I felt that the timing could be improved.

The timing of the gestures was great none of them was noticeably out of sync with what was being said.

Judges were also asked to comment on what issues with timing might be improved. No major issues were revealed.

Homonym	Gesture Type	Gesture
Hear-Here	Iconic	Hold hand near ear
Hear-**Here**	Deictic	Move hand toward torso
Moose-Mousse	Iconic	Hold hands up like antlers
Ate-Eight	Iconic	Move hand toward mouth
You-Ewe	Deictic	Point hand away
Him-Hymn	Deictic	Point hand away
Hoarse-**Hoarse**	Iconic/Deictic	Hold hand near neck
Feat-**Feet**	Iconic/Deictic	Point hand toward feet
Sea-**See**	Iconic	Hold hands above eyes
Whale-**Wail**	Iconic	Rotate hands near eyes
Paws-Pause	Iconic/Deictic	Hold hands near torso
Hair-Hare	Iconic/Deictic	Hold hands near head

Table 3: Gesture dictionary

2.4 Recruitment, Experimental Procedures, and Data Analysis

We recruited participants from two different classrooms within the Children's Learning Center of UW-Milwaukee, a large public university in the Midwestern United States. Parental consent forms were distributed to parents of all enrolled children between the ages of 4 and 8 before conducting the research study. For each child to participate, the parent of a child was required to sign and return the permission letter and the child had to give his or her own verbal consent.

Children in the study were randomly assigned to one of two groups: either the control (without gestures) or the experimental condition (with gestures). The robot performed the story to the two groups separately on two consecutive Friday afternoons, during their regularly scheduled time at the Children's Center.

Children in both groups were first given a pretest and afterward a post-test, while sitting at small tables of three to four people with a staff member at each table to help them. The pretests and post-tests both comprised the same set of ten items each of which included two full-color cartoon-style images, selected to depict two distinct words that are homophones, similar to the illustrations used in published versions of the story. Table 4 includes the sentences and a brief description of the images from which the children could choose. The children were read the sentences by an adult and asked to circle the image that they felt best corresponded to the meaning of the sentence.

After the pretest, the children moved to a nearby area where they were were seated on the floor near the robot. There, they were introduced to the researcher and the robot. Then they heard the robot tell the story which took about three minutes. After the story, the children returned to the tables with the caregivers for the post-tests.

Precision was measured for each condition by calculating the number of correct responses divided by the total number of responses for each condition, aggregated over the entire group and the complete set of items. To account for possible differences between the two groups besides the condition, we calculated two measures of significance: one between the post-test values of precision for the control and experimental groups and one between the pretest and post-test values within the experimental group alone. We also looked at the differences between the number of incorrect versus

Sentence	**Pair of Images**
1. The sea was as smooth as glass.	ocean waves, eye
2. He ate twice the amount that you did.	boy eating sandwich, number eight
3. She didn't come here to talk to me.	girl pointing to floor, hand cupping ear
4. She felt the hair rising on the back of her neck.	hair without face, rabbit
5. Hey, where are you going?	boy waving, bale of yellow hay
6. A whale can be very loud underwater.	boy with tears, whale spouting water
7. We had some mousse for dessert.	walking moose, dish of swirled brown food
8. You again?	generic sheep, hand pointing towards reader
9. The teacher's criticisms were enough to make anyone red.	girl reading book, splat of red color
10. I was driving along the road when a deer jumped out at me out of the blue.	boy blowing candles, splat of blue color

Table 4: Test sentences with descriptions of image choices

correct responses for individual items, aggregated over the entire group. Any response that had two circles was counted as incorrect based on the assumption that the child was unable to disambiguate the lexical ambiguity of the sentence.

3. Results

Twenty-two adults from a university course on Natural Language Processing viewed a video of the robot storytelling and assessed the timing of the gestures. Most (18 of 22) rated the timing of the gestures as natural ("great": 14; "good": 4 ; "okay": 2; "poor": 2). Among those who gave it a lower rating, the duration of two of the gestures was cited, including two who said pointing to the feet was too slow and three who said pointing to the eyes was too quick. However, no major problems with the gestures was revealed.

The children in the homophone study ranged from ages three to seven. Subjects in the control group had ages ranging from three to five, while subjects in the experimental group had ages ranging from four to seven. 13 students were tested in the control group pretest and 12 students were tested in the control group post-test. Thus the aggregated number of responses for the control pretest was 130 and for the control post-test it was 120. 8 students were tested in the experimental group, i.e. the group with gesture, both pretest and post-test. Thus for both the pretest and post-test there were 80 responses.

Table 5 shows the pretest and post-test precision for both the control (no gesture) and experimental (with gestures added) condition. The total number of correct over the total number of responses is shown in parentheses.

Pretest CON	**Post-test** CON	**Pretest** EXP	**Post-test** EXP
0.762 (99/130)	0.633 (76/120)	0.875 (70/80)	0.925 (74/80)

Table 5: Pretest and post-test precision for control (CON) and experimental (EXP) conditions

In the control group, aggregated precision dropped by 0.109, which is 14.3 percent; in the experimental group, the aggregated precision increased by 0.050, which is 5.7 percent. The difference in aggregated precision between the control and experimental groups for the post-test is significant ($P < .001$). The increase in aggregated precision within the experimental group is promising, but insufficient to reject the null hypothesis ($.1 < P < .2$). Looking at individual items, for the control group, the number of correct responses increased for two items (Here-Hear and Red-Read), stayed the same for two items (Eight-Ate and Moose-Mousse) and decreased for 6 items. By contrast, for the experimental group, the number of correct responses increased for 4 items, stayed the same for 6 items, and never decreased. (Charts capturing the complete results of all the pretests and post-tests for both conditions are provided in Appendix A.)

In the control group (hearing the story without gestures), there were several homophones that a majority of the children found difficult. In the pretest, children struggled most with Moose-Mousse, which was the only pair that had more incorrect responses than correct responses (9 incorrect, 4 correct), followed by Read-Red (6 incorrect, 7 correct), and then Here-Hear (4 incorrect, 9 correct). (None of these pairs involves a syntactically ambiguous word; for the target audience both “moose” and “mousse” might be rare.) The children appeared to excel at understanding Wail-Whale (100 percent of the students got this correct), Sea-See, and Blue-Blew. Overall, the children in the pretest control group had 99 correct responses and 31 incorrect responses. The control group post-test had significantly different results. Children still struggled with Moose-Mousse (9 incorrect, 3 correct). The results showed improvement among the homophone Here-Hear (22 percent increase), but a slight decrease in most other homophone pairs. In the post-test control group, there were 77 correct responses and 43 incorrect responses.

In the pretest of the experimental group, children had some difficulty with Here-Hear, Hey-Hay, and Mouse-Mousse. Despite having some mistakes, however, none of the pairs were chosen incorrectly more than correctly and none involves a syntactic ambiguity. There were a total of 6 incorrect homophone pairs (Sea-See, Eight-Ate, Here-Hear, Hey-Hay, Moose-Mousse, and Read-Red), with 10 incorrect responses total. The children did have a few perfect scores as well for Hair-Hare, Ewe-You, Blue-Blew, and Whale-Wail. In the post-test, the experimental group showed more uniform improvement. There was an overall decrease in the number of incorrect homophone pairs, which was 4 pairs and 6 incorrect responses total out of 80 responses. The children had 100 percent accuracy for six of the ten items: Eight-Ate, Hair-Hare, Whale-Wail, Ewe-You, Read-Red, and Blew-Blue. Two of the items answered correctly in the post-test had been among those that were incorrect in the pretest (Eight-Ate, Read-Red); neither involves a syntactic ambiguity. Pairs that were still sometimes resolved incorrectly involved words that might be more unusual and also harder to gesture (Hey-Hay, Here-Hear, Moose-Mousse); only one was not also discriminated by the syntactic context (Moose-Mousse).

4. Discussion

The results of the study suggest that adding coordinated gestures can have a positive impact on children’s ability to distinguish between homophones in the context of robot storytelling, and potentially more broadly. Our work also suggests that studying competency in understanding homophones is an interesting measure of human language development.

As we noted earlier, much past work with children focuses on children’s own use of gestures to discriminate meaning (see Kidd and Holler, 2009; Ray, 2015), which would suggest that children can understand gestures at young ages. Studies directly related to children's perceptions of gesture have considered either vocabulary acquisition or object labeling. Vocabulary acquisition among infants (ages 9 to 15 months), measured as the number of new words produced or recalled, has been found to be larger when mothers use iconic gestures versus uncoordinated gestures, e.g. (Zammit and Schafer, 2011). Recall of object labels among infants is also improved by parents' use of iconic gestures (Namy et al., 2000). Similarly, the work by Kory Westlund et al. (2017) found that among preschoolers (average age of five years) vocabulary acquisition was better when social robots used gestures. None of this prior work has directly addressed the challenge of understanding homophones.

Our results reveal a clear benefit to gestures in homophone discrimination. Precision in the experimental group, which saw the gestures, increased from an already high value of 0.875 to a near-perfect precision of 0.925. For only two items (Here-Hear and Hay-Hey) did more than one child make a mistake on the post-test. This improvement is in sharp contrast to the results in the control group where the responses seemed somewhat random and actually declined from pretest to post-test. The only item where the performance of the control group improved was on arguably the hardest item (Here-Hear), which had the most errors in the pretest of the experimental group.

We did not make any adjustments to either the pitch or intensity of the homophones, as it would not be possible to do so uniformly given the sometimes high density of homophones within a sentence. However, in the future we would like to look at increasing the pitch or intensity of homophones that are not surrounded by other homophones, especially those that occur at the beginning or end of syntactic phrases. This method of emphasis seems to be used by human storytellers reading *Dear Deer*. Also the Theune et al. (2006) model suggests rules for these, although the researchers themselves did not evaluate their impact at the time due to limitations in their text-to-speech software.

We acknowledge several limitations of the study. The number of children in both groups was small. Also, the number in the experimental group was smaller than in the control. We found predicting attendance ahead of time can be difficult, as attendance at a university-based daycare center can vary significantly, due to schoolwork or travel, which affects both student and faculty parents. We suspect that having a larger number of children in the control group, including more children who were at the low end of the age range and fewer at the high end, likely contributed to some of the variability we saw in the pretest to post-test results for the control versus the experimental group.

The children in both groups were relatively young and this factor, along with possible excitement or fatigue, might have affected their ability to understand or follow instructions. To reduce this risk, the children were seated at different tables in groups of around four, with a staff person at each table to explain the procedure and remind the children not to discuss their answers. We did observe, however, that occasionally the children would not comply and had to be reminded. Thus, while some sharing of answers might have occurred, it was likely limited to a few items, and only among children within a table group.

The results for any particular pair of homophones may have been influenced by a number of potential linguistic effects that would occur when reading an existing story to children. The homophones might vary in their difficulty for children, even more than for adults, either because some words might be more or less familiar to an individual child (i.e. people's lexicons grow as they interact with language) or because the words vary in their linguistic features (e.g. some are nouns, some are verbs, some can be both). The homophones chosen to be gestured were based on the feasibility of coming up with a coordinated gesture given the limitations of the robot and the phrasing of the story. The test sentences were chosen such that only one of the answers would be semantically appropriate; in most cases (6 out of 10) only one answer would have correct syntax, based on judgment of an adult native speaker of English, with expertise in linguistics (the second author). Thus, the linguistic context surrounding each homophone (either in the story or in the test sentences) could also potentially have had an impact. Whether a given child is aware of any of these aspects is something that would be hard to predict, however, as children's ability to recognize linguistic features has been found to improve as they mature (Clahsen et al., 2007). Measuring understanding individually and dynamically, such as through event-related brain potentials (ERP), might be interesting, but it would be a very different study. Such measurement would also not contribute to the motivating goal of our work, which is to see if adding gestures to a humanoid robot reading a story in a classroom setting would be educationally worthwhile. That goal was met, as the change in performance from pretest to post-test was much better in the experimental group than in the control.

The results on the pretests were very different between the control and the experimental groups, with the experimental group having higher overall precision on the pretest. This difference might be explained by the natural variability in ability among new language learners or by differences in the distribution of ages within the two groups. We observed a slightly higher proportion of older children (more six or seven year olds) in the experimental group than in the control. This could be a factor, as prior research suggests that some children's ability to detect lexical ambiguity does not emerge until around six years of age, as reported by (Kidd and Holler, 2009); however they did find some subjects as young as three could understand and perform gestures appropriately.

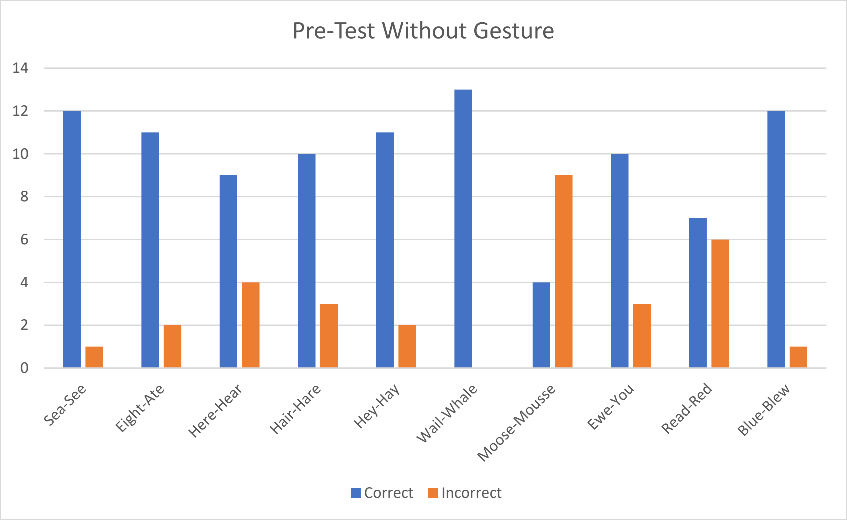
Some of these factors, such as the age of the children, are unavoidable if one wants to study children's language development. However, in a future study, one could do a similar experiment using ESL learners, as these subjects would be more mature and less prone to outside distraction. Some of the limitations could be addressed by conducting a broader study with assignments to groups conditioned on the results of pretests conducted ahead of time.

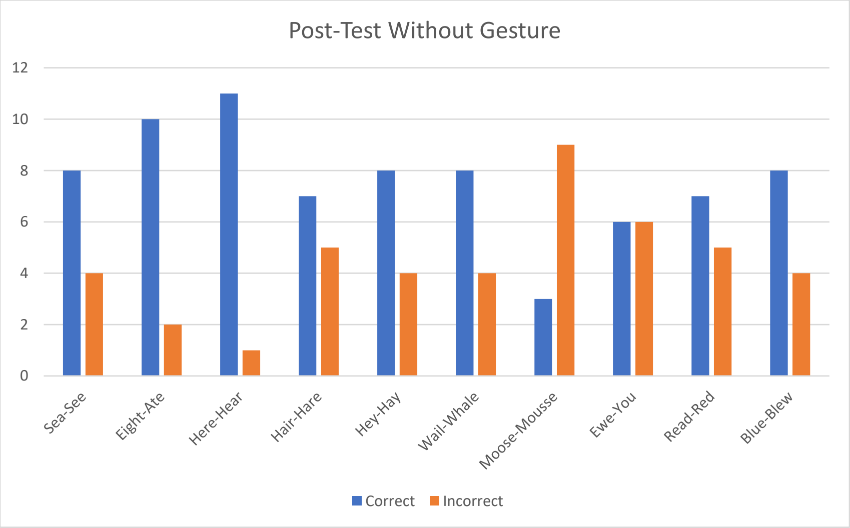
Another aspect to explore would be the types of gestures used. Using humanoid robots offers the opportunity to control exactly what gestures are produced when, and to assure they are repeated exactly the same way each time. Currently, we do not know if the semantics of iconic and deictic gestures is more important than just the fact that using multiple modalities, in an apparently coordinated way, creates an emotional response that enhances learning and memory. It may also help to keep young children's attention to other supporting details. In a future study, it would be interesting to compare the effect of semantic gestures to the use of coordinated, but simpler, gestures. Simple beat gestures, such as moving arms up, down, or forward, in time with the phrasing, might be sufficient to enhance engagement. (One of the judges who viewed our video of the robot gestures reported that it seemed unnatural when the robot was not using any gestures.) If coordinated beat gestures are enough, then, in addition to assessing the impact on understanding homophones, one could explore how robot gesture affects the retention of narrative details, using a story that has distinct characters, setting, or events.

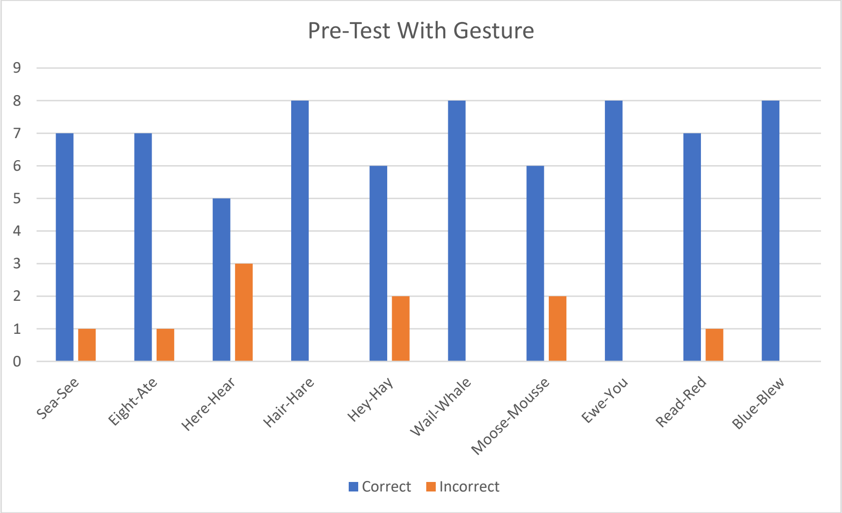
5. Conclusion

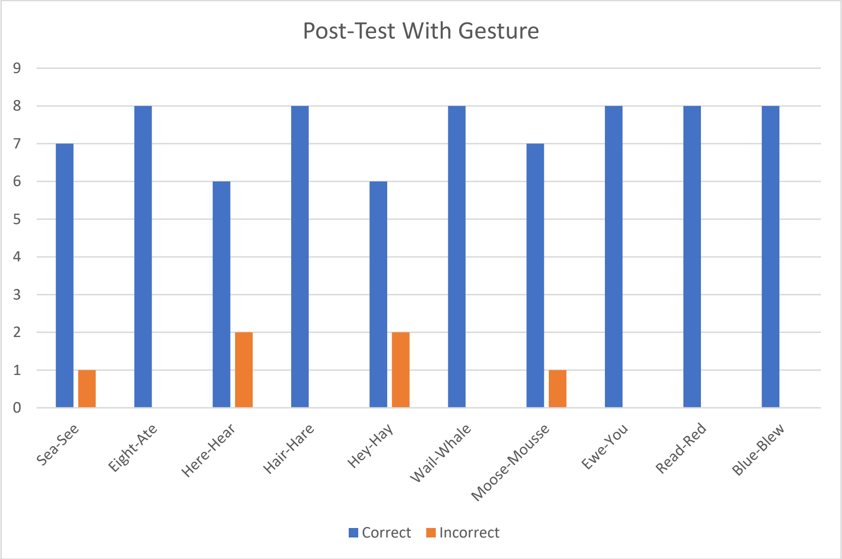
This paper describes an experiment conducted in the context of human-robot interaction, to assess the impact of iconic and deictic gestures on the understanding of homophones by young children. We used a humanoid robot and its associated software framework to present a short story containing pairs of homophones to small groups of children, accompanied by either these expressive gestures or no gestures. What we found is that the use of iconic and deictic gestures appears to aid these language learners in their general understanding of homophones. This work thus provides valuable additional evidence for the importance of gesture to the development of children's language and communication skills.

APPENDIX A: Pretest and post-test results for reading study









ACKNOWLEDGMENTS

The authors thank the staff and families of UW-Milwaukee Children's Learning Center for the use of their facility and their assistance and enthusiastic support for the study.

References

Gene Barretta. *Dear Deer: a Book of Homophones*. Square Fish, 2010. A previous hard copy edition was published by Holt & company in 2007.

Harald Clahsen, Monika Lück, and Anja Hahne. How children process over-regularizations: Evidence from event-related brain potentials. *Journal of Child Language*, 34(3):601–622, 2007. doi: 10.1017/S0305000907008082.

Paul Ekman and Wallace V. Friesen. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1(1), 1969.

Jana M. Iverson, Olga Capirci, and M. Cristina Caselli. From communication to language in two modalities. *Cognitive Development*, 9(1):23 – 43, 1994. ISSN 0885-2014. doi: 10.1016/0885-2014(94)90018-3.

Angelina Jean. Dear Deer, By: Gene Barretta, Read By: Angelina Jean, 2015. URL `https://www.youtube.com/watch?v=jYaMP36D8s0`. Accessed August 2, 2018.

Evan Kidd and Judith Holler. Children's use of gesture to resolve lexical ambiguity. *Developmental Science*, 12(6):903–913, 2009.

Jacqueline M. Kory Westlund, Sooyeon Jeong, Hae W. Park, Samuel Ronfard, Aradhana Adhikari, Paul L. Harris, David DeSteno, and Cynthia L. Breazeal. Flat vs. expressive storytelling: Young children's learning and retention of a social robot's narrative. *Frontiers in Human Neuroscience*, 11:295, 2017. ISSN 1662-5161. doi: 10.3389/fnhum.2017.00295.

David McNeill. *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press, Chicago, IL, 1992.

Laura Namy, Linda Acredolo, and Susan Goodwyn. Verbal labels and gestural routines in parental communication with young children. *Journal of Nonverbal Behavior*, 24(2):63–79, 2000.

Elizabeth M. Ray. Gestures used by ESL children to resolve lexical ambiguity. Master's thesis, Ohio University, 2015.

TeachingBooks.net. Website, 2018. URL `https://www.teachingbooks.net/tb.cgi?tid=16401`. Search key: Dear Deer. Accessed August 2, 2018.

Mariët Theune, Koen Meijs, and Dirk Heylen. Generating expressive speech for storytelling applications. *IEEE Transactions on Audio, Speech and Language Processing*, 14(4):1137–1144, 2006.

Elisabeth T. van Dijk, Elena Torta, and Raymond H. Cuijpers. Effects of eye contact and iconic gestures on message retention in human-robot interaction. *International Journal of Social Robotics*, 5(4):491–501, Nov 2013. ISSN 1875-4805. doi: 10.1007/s12369-013-0214-y.

Maria Zammit and Graham Schafer. Maternal label and gesture use affects acquisition of specific object names. *J Child Lang*, 38(1):201–221, 2011. doi: 10.1017/S0305000909990328.