Robot learning of lexical semantics from sensorimotor interaction and the unrestricted speech of human tutors

Joe Saunders, Chrystopher L. Nehaniv and **Caroline Lyon**[1]

Abstract. This paper describes a HRI case study which demonstrates how a humanoid robot can use simple heuristics to acquire and use vocabulary in the context of being shown a series of shapes presented to it by a human and how the interaction style of the human changes as the robot learns and expresses its learning through speech. The case study is based on findings on how adults use child-directed speech when socially interacting with infants. The results indicate that humans are generally willing to engage with a robot in a similar manner to their engagement with a human infant and use similar styles of interaction varying as the shared understanding between them becomes more apparent. The case study also demonstrates that a rudimentary form of shared intentional reference can sufficiently bias the learning procedure. As a result, the robot associates human-taught lexical items for a series of presented shapes with its own sensorimotor experience, and is able to utter these words, acquired from the particular tutor, appropriately in an interactive, embodied context exhibiting apparent reference and discrimination.

1 INTRODUCTION

Humans acquire language quite effortlessly during their early years and continue to refine and develop their language ability throughout their lives. There are many alternate views on how language is acquired varying, for example, from the Chomskyian [4] idea of existing innate pre-cursors of syntactic structures and mechanisms which support the acquisition process to alternatives suggesting that that no 'specialist' functions exist and where more general learning mechanisms, including learning from social interaction, are used [24]. In our work we explore this latter view and consider that language learning is a social and interactive process. We review research on how language is acquired in human infants and suggest that similar mechanisms might be applied to robots in order for them to acquire linguistic capabilities. In fact we believe that it should be possible for a robot to use what it learns individually and from others socially through grounded sensorimotor interaction to bootstrap the acquisition of language which in turn will create a positive feedback cycle between using language and developing other cognitive abilities.

In this paper we report on an initial case study where, through unrestricted speech and interaction with a human partner, a robot can attach meaning to a series of individual shapes. To do this we firstly consider how human infants acquire such meanings and then carry out experiments to see whether broadly equivalent mechanisms will allow a robot to do the same. Two research questions are explored, firstly, will a human interaction partner employ similar and effective interaction styles (such as child directed speech) with a robot partner?, and secondly, could a robot be able to extract sufficient information from such interaction episodes in order to ground the meaning[2] of the lexical items (words) used when taught by particular humans about a set of physical shapes?

2 BACKGROUND

In this research we take a 'usage-based' view on language acquisition following Tomasello [24] and Bloom [3] and take inspiration from the constructivist work of researchers such as Steels on the emergence of various linguistic capacities in agent communities [21], and also Roy and Pentland [18], Roy [17], and Yu and Ballard [27] in attempting to link perceived speech with object and action perception. We differ from the above in that the focus of this research considers the effect of the robot feedback in modifying the interaction style of the human tutor and how this change enhances (or worsens) the learning experience of the robot.

Social interaction appears to be essential for effective language acquisition is children [19]. Adults do not specifically attempt to 'teach' language to children but rather aim for joint understanding with the child. This often comes about by the adult using a form of language called Child Directed Speech (CDS) which is specifically tailored to the perceived level of linguistic skill of the child [5]. CDS differs from adult directed speech in that adults speak more slowly, often repeat what is said and may correct mistakes by the child during the interaction as well as prominently drawing attention to topics via changes in pitch, pause duration, word placement, word duration and intonation. As joint understanding between the child and adult is perceived by the adult to grow, fewer aspects of CDS are employed. However, although infants are exposed to a high number of ambient speech events, being exposed to as many as 7,000 utterances per day [3], they actually learn new words with very little exposure. According to Bloom [3] the natural co-occurrence of seeing an object and hearing the object name does not occur regularly enough to allow an associative mechanism to differentiate the object. In order to achieve learning in few episodes the infant must have its learning experiences biased in some way. This is thought to occur via intentional reference. Thus for example a 9-month old baby will follow the gaze of its mother, follow her pointing gestures and monitor her emotional states. By 1 year old, the infant points on its own and observes the adult's gaze whilst checking if it has changed the adult's attention. If the child fails to capture the adult's attention it will alternate between gazing at the object and the adult until it succeeds in getting the

[1] Adaptive Systems Research Group, University of Hertfordshire, College Lane, Hatfield, Herts, AL10 9AB, UK, email: j.1.saunders@herts.ac.uk

[2] Following Wittgenstein [26], any derivation of 'meaning' must ultimately be evidenced by appropriate embodied action in language games (cf. [12]).

attention of the adult onto the object. The utterances of the adult together with reinforcement via forms of affective feedback (prosodic features of CDS) and success/failure of shared intentional reference for the infant whilst situated in context allow enough bias for fast learning to take place. Within this shared context the action modalities of the child are associated with simple verbal speech patterns of the adult. In our experiments learning relies on the association of perceived speech and action modalities of the robot together with the robot's ability to share context via a form of rudimentary referential 'intent' with a human teacher.

In order for our robot to be attentive to the topic spoken by the human we currently only consider *pause duration* and *word placement and word duration*, although in future research we will also consider prosody via changes in pitch and intonation. In terms of word placement and duration we make the assumption that regularities can be extracted from the human's speech stream. In English (and other languages) attention is drawn to new nouns by placing them at the end of an utterance (reflecting the subject-verb-object structure of the language) [5, pages 62-66]. There is also evidence that a similar phenomenon occurs in CDS in other languages (e.g. Turkish) even when to do so would appear to be ungrammatical or uncommon in adult speech for that language [2]. As well as indicating salience via utterance-final word placement, often the word is also spoken with a longer duration than average during the utterance. We therefore use both the duration and placement of a word in the perceived speech matched with the sensorimotor stream experienced by the robot in order to bias the robot's learning ability.

Clearly it would be trivial to provide the robot with a set of pre-programmed rules attaching a label to a given perceived object, however our research attempts to allow the robot to acquire lexical semantics ('meaning') of words used by a human tutor for the objects (in terms of the set of sensory-motor experiences over time) dynamically through the interaction process. We believe this is an important step not only towards the robot understanding objects in its world (the focus of this study) but also understanding actions and states (e.g. 'push the star','the star is next to the square'). This latter step forms part of our future research and which we hope will provide entry into basic grammatical constructions [24].

3 RESEARCH QUESTIONS

In this study we considered two issues, the first from a human-robot interaction perspective and the second from the perspective of learning efficiency (in terms of classification performance) of the robot during the sessions. Exploring the first issue, we expected that the human interaction partner would employ a language style similar to that employed in child directed speech and that with each subsequent interaction session the feedback from the robot (the robot in effect expressing what it had learnt during the previous sessions) would motivate the human partner to modify the interaction, supplying greater or lesser focus as the abilities of the robot improved. Note that although the participants were specifically asked to treat the robot as a 1-2 year old child (see 4.1 below) and the robot had the appearance of a young child, it was by no means certain that the interaction style would 'mimic' CDS as other factors which may provoke CDS with human children (such as emotional attachment and richness of interaction) may be missing from the interaction with the robot. Thus the first two research questions were:

1. Does a human interaction partner engage with a humanoid robot in a verbal style similar to that of Child Directed Speech?

2. Does a human partner modify aspects of their robot-directed speech as the robot learns and changes its interaction capability?

The second issue explored was that of the robot's capacity to learn, given the interaction style of the human. We expected that the robot would become more skilled as the human partner interacted with it which may in turn motivate the human. This skill level of the robot indicating that the multi-dimensional set of sensorimotor attributes had become refined such that relevant meaning emerged. This latter aspect being measured by a number of factors, including shape classification performance and increasing mutual information of particular sensorimotor attributes. Thus the final research questions explored were:

3. Does a robot's ability to recognize and correctly classify shapes by its utterances improve during the interaction sessions?

4. Can a robot select those parts of the sensorimotor stream that are relevant to classifying presented shapes by acquired vocabulary?

4 CASE STUDY DESCRIPTION

In this section we describe both in summary and in subsequent detail the experimental procedures in the case study.

4.1 Experimental Summary

In this experiment we asked participants (see 4.3) to explain/teach (without any constraints on their language) a series of shapes (see 4.5) to a humanoid robot during a series of interactions *as if* the robot were a 1-2 year old child (see figure 2). The robot was pre-programmed to track and habituate for a given period on these shapes. Following each interaction the speech stream of the human was converted into phoneme strings marked with word boundaries. These phoneme strings were subsequently aligned with the sensorimotor modalities experienced by the robot during the interaction session. The aligned speech and sensory modalities were then processed to highlight words of long duration and that appeared at the end of utterances. This processed modality stream became the basis for the robot's learnt experiences for the next interaction session with the human. In these subsequent sessions (from sessions 2 onwards) the robot was allowed to match its current sensorimotor input (that it was experiencing during the interaction) against that learnt in the previous session(s). This allowed the robot to react to the human by expressing (via its own speech) what it had learnt during the previous session(s). Thus, for example, when presenting the 'moon' shape and depending on how the human had previously described it, the robot might say 'moon'. Alternatively if the human had used different words (or different pronunciations) the robot might say for example 'mooone', 'crescent' or 'smile' etc.

4.2 Software Platform and Architecture

In carrying out these studies we have used an existing social learning architecture (ROSSUM [20]). The architecture has been used to learn scaffolded behaviours via directed learning from a human teacher (but not previously linguistic behaviour). ROSSUM allows experiments to be carried out where, via the processes of self-imitation and observational imitation, a robot learns through experiences grounded[3] on its own visual, auditory and sensorimotor feedback and the relevant interaction modalities presented by a human

[3] The idea of 'grounding' refers to relating the meaning of symbols to embodied sensorimotor experience [9].

tutor. In this experiment *speaking* is regarded as a particular type of *gesturing*, i.e. motor activity for manipulating the physical or social environment (see [13, 15]). Actions which can be taught to the robot (see figure 1 below) now include speech acts (in this case uttering words represented as phonemic strings in a learnt experiential context). These modalities are in addition to existing object detection and proprioceptive feedback modalities already present in the ROSSUM architecture (see 4.4).

4.3 Participants

Eight adult participants took part in the case study. All participants were between the ages of 27 and 58 comprising 5 females and 3 males. The backgrounds of the participants were either administrative (6) or research related (2), the latter not connected with robotic language research. Four of the female participants cared for either children or grandchildren under the age of 5.

Each of the eight participants took part in 5 interaction sessions of approximately two minutes with the robot (in total 40 robotic interaction sessions) and all of the sessions were videotaped for later analysis. Prior to the interaction sessions each participant also had 4 training sessions with an automatic speech recognition system. Unfortunately initial results from using this system were very poor and an alternative was subsequently used (see 4.6 below).

The experiment was carried out over a three month period between March and June 2009 based on availability of the participants. Participants were paid a small stipend of £20 if they completed all 9 sessions (4 training + 5 interaction) and all participants did complete all sessions.

4.4 Robot

The experiment used the Kaspar2 robot (see figure 1). This is a minimally expressive small humanoid robot designed for HRI studies [7] with 8 degrees of freedom in the head, and 5 degrees of freedom in each arm (although the arms were not used in this experiment).

Figure 1. ***Learning via Social Interaction.*** *Here (in a previous experiment) the humanoid robot Kaspar2 is taught how to hold a patterned box via interaction (manipulation of its joints) with a human teacher.*

Kaspar2 employs two video cameras for its eyes. These are used to obtain images which are processed [14] to yield an additional set of modalities including object/pattern recognition and face detection. In this experiment the following set of sensory-motor modalities were recorded by the robot at approximately 30ms intervals - face detection (binary), object id, head pan, head tilt, head roll, position of object in visual field (x,y position of centre of object) and distance to object.

4.5 Shape recognition

Six graphic shapes were used and pasted onto the sides of a box (a square, circles, triangle, sun, moon and star - note however that no verbal prompting or description of these shapes was given to the participant). These objects are detected using the ARToolKit system [1]. The objects are pre-learnt using ARToolKit, thus Kaspar2 can detect these objects and recognize that they are individual entities in the world (effectively encoded by the integers 1-6 in a sensory stream), however *no other 'meaning'* is attached to them. We justify the use of this simplifying step in these experiments as, firstly, it eliminates the need for a complex vision processing modality and secondly, and more importantly, it reflects a 'whole object' bias found both in children and adults (see Bloom, Chapter 4 [3] and Mervis [10]).

Figure 2. ***Sharing Reference with a Teacher in Context.*** *Kaspar2 interacts with the teacher, where both the robot and the human share attention on the shapes on the box, a case of rudimentary shared intentional reference.*

4.6 Speech Processing

Initially a standard speech recognition system [11] was used to process the speech stream of the human. However it was found that the accuracy of such a system rapidly decreased when dealing with the unrestricted speech stream used when interacting and teaching the robot. An alternative approach was taken which used speech alignment software (see [22]). This software takes as input a transcript of the human's speech during the interaction session together with the recorded speech stream (as a .wav file). It outputs timings, phonemes and word boundaries for the recorded speech.

4.7 Rudimentary Shared Intentional Reference

In the experiment it was important that the robot and human share attention and focus on the same objects whilst simultaneously provoking actions which would motivate the human teacher to explain the shapes. We make a key assumption - that the robot is motivated by novelty. This was achieved by having Kaspar2 fixate and track shapes as they entered its camera images. When a shape was not present Kaspar2 would search the environment semi-randomly (randomly within fixed limits) by varying the control signal sent to its pan/tilt/roll head motors. If a shape was found Kaspar2 would 'smile' and fixate on that object. A set of linear 'boredom' filters, one for each shape and for the human's face, would saturate if the same shape (including the face) remained in focus for more than $\sim$20 seconds. A saturated filter would cause the robot to begin searching for a new object (and Kaspar2 would stop smiling). The filters would linearly

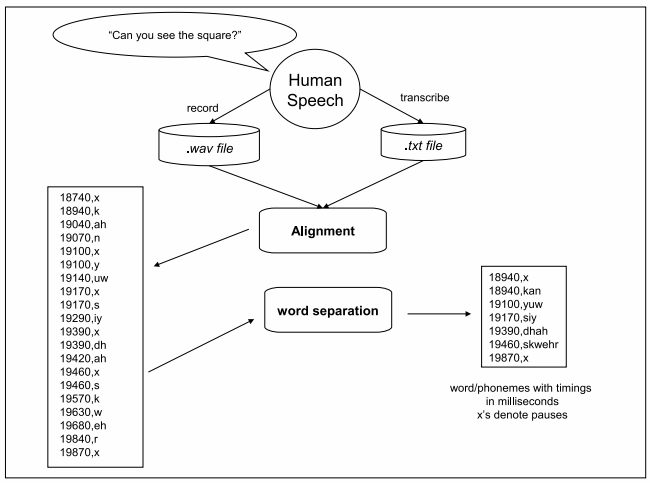

Figure 3. ***Segment of Words/Phonemes output.*** *The human's speech stream is converted to a file of timed phoneme/word outputs (x's denote word boundaries). Note that different speech patterns can produce different phoneme strings for the same word e.g. different vs. diffrent (i.e. middle 'e' not expressed).*

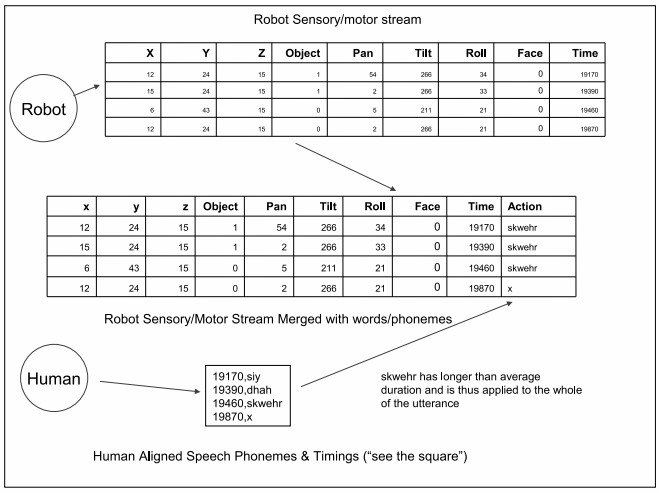

Figure 4. ***Aligning Human Speech Words/Phonemes with Robot Sensorimotor Streams.*** *The timed word/phoneme strings are aligned with the robot's sensory-motor stream based on time. Here the word 'square' is an end of utterance word with higher than average duration and so is propagated over the whole of the preceding utterance and concommitant sensorimotor experience. This results in a table subsequently used for sensorimotor similarity matching (using* k*-Nearest Neighbour) in the robot's execution phase. The 'action' executed in this case would be a robot speech output (i.e. if similar to the current sensorimotor state the word 'square' would be expressed by the robot).*

de-saturate the longer the object was not in focus. The net effect of these measures was for Kaspar2 to appear to be interested in the objects/face but become rapidly bored if viewed for too long. This had the effect of provoking the human teacher to try to gain the attention of the robot with new shapes and subsequently explain them.

4.8 Attaching Meaning to Words for Shapes

In this context we consider that 'meaning' of a communicatively successful utterance is grounded in its usage based on the robots sensorimotor history from acting and interacting in the world. These grounded meanings can then be scaffolded via regularities in the recognized phoneme/sensory-motor stream of the robot. The first step in this process is to merge the speech stream of the human (represented as a set of phonemes with word boundaries) with the robot's sensorymotor stream. This is achieved by matching the two modalities based on time (see figure 4).

This merger effectively associates what was said to the robot with what was experienced by the robot at that time. Within this study the set of sensorimotor attributes is limited to the robot's head and vision proprioception together with the recognition of pre-trained shapes. However although the categorization of the shapes is already available (as a integer) the robot has not associated this attribute with either the words in the human's speech stream or its own proprioception. In this study sensorimotor attributes which do not affect the primary association of object and word (such as the head and image proprioception) should become less relevant over time (i.e. a 'moon' shape is still a moon shape no matter where or how it is seen). To achieve such associations we face a number of difficult challenges. Firstly is that of associating what was said to the appropriate parts of the sensorimotor stream. Thus the human tutor may show the robot a shape (e.g. the 'triangle'), but only say the word 'triangle' within the utterance before or after the shape has appeared/disappeared from the view of the robot (e.g. 'here's a triangle' and then show the triangle or 'that was a triangle' after having shown the triangle). Secondly, which of the set of sensorimotor attributes and at which points in time are such attributes relevant to the speech act? We make no pre-programmed choices as to what is relevant for the robot however in the simple scenario presented here the most important will be the object identifier returned from the ARToolKit system. But we have not only the object identifier but also the modalities from the camera image (x,y,z dimensions of the object in space), the encoder readings from the robot's pan/tilt/roll head unit and the binary value of the face detection algorithm. What makes any of these less (or more) relevant than any other? In order to cope with these issues we apply two heuristics. Firstly, that the relevant part of the human speech stream is the word expressed at the end of an utterance and with a higher than average duration (following the ideas discussed in section 2 above). We extract a word if it occurs at the end of an utterance (the end of an utterance defined as a pause longer than average word duration). This word is then remapped back onto each element of the time sensorimotor stream of the previous utterance. In effect this makes the word chosen relevant to the whole of that utterance and thus relevant to any sensorimotor inputs which arose during that time. This means that most (though not all) of the timing effects mentioned above can be avoided as the relevant word will occur thoughout the utterance. Secondly, having associated the relevant word with the utterance/sensorimotor stream we then compute the *information gain* [16] between the sensory attributes and the chosen word (effectively a measure of *mutual information* indicating the expected amount of information (in bits) that discriminates the given word by the given attribute). This measure is used during the next interaction sessions to weight the similarity measure of current vs. stored experiences. This is described in section 4.9 below. Note, in our experiments, the robot learnt lexical semantics separately from each participant so that learning occurred in effect as if each partic-

ipant had their own robot that had learnt only from them. This allowed us to analyse the diversity of learning trajectories, dependent on interactions with particular tutors (who could in principle be using completely different lexicons or even different languages).

4.9 Execution and Action Selection

We consider robot speech acts (including 'saying' nothing) to be no different from any other 'action' (such as primitive head or arm movements) that may be selected for execution. Therefore an existing selection method, based on matching the current sensory stream against a set of pre-learnt/pre-taught behaviours, is used [20]. All previously learnt action/state sequences are held in *memory models*. These are tables in which each individual row is a state/action pair. Actions can refer to individual primitives or other memory models. Execution starts at the topmost memory model and recurses though the possible set of models until a primitive (or marked goal state) is found. For the language experiment discussed here only one memory model is used. This equivalently contains the set of state/action pairs learnt by the robot (see 4.8 above). Here the 'action' is uttering a phoneme string representing a word. The phoneme strings are output by the robot using the eSpeak speech synthesizer [8].

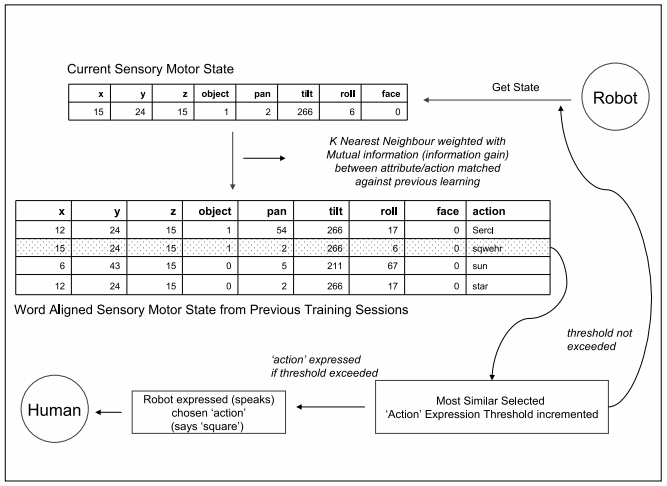

Figure 5. ***Interaction Execution Loop.*** *The robot continuously monitors its state. This is matched with the table previously learnt using the* k*-Nearest Neighbour algorithm. Dimensional attributes are weighted using information gain. The nearest neighbour is selected. The 'action' part of the selected previous learning will only be expressed (spoken) by the robot if it has been chosen a given number of times (exceeds a predefined threshold). Note that the execution loop and the learning loop (where the robot's proprioceptive state from the current session is recorded) happen simultaneously.*

The robot continually polls its current state (see figure 5). This state is matched against the memory model using the *k*-Nearest Neighbour (*k*NN) algorithm with dimensional attributes weighted using information gain. We use the TimBL memory based-processing system to carry this out [6, 23]. In this experiment the *k* value was set to 1 and similarity metric used is a 1-norm Manhattan distance. The *k*NN algorithm always returns a classification, in this case an 'action'. This is effectively a prediction based on matching what the robot experiences and what the robot should then 'say'. The robot state changes rapidly during the interaction and therefore rather than simply express each matched state the action (a phoneme/word string) will only be expressed once it exceeds a given threshold. The selected action is incremented if matched and all other actions penalized. If the same action is predicted successively it receives a bonus which serves to reward consistent prediction (for complete details of this scheme see [20]). The net effect is that by lowering the threshold more matches may be expressed by the robot (and thus similar words matching the same state may be output (e.g. participants might have described the 'square' as both a 'square' and a 'box', the robot may choose either). Raising the threshold means that fewer words are expressed and the similarity criteria to the current set of states is higher (e.g. only 'square' would ever be expressed). In this experiment words are output (including silence) with a threshold, reward, penalty and bonus set during pre-experimental tests providing a balance between an overly talkative robot and one who said very little.

5 RESULTS AND ANALYSIS

As discussed in section 2 above, CDS is characterized by slower speech, repetition, mistake correction, prosody, duration of pauses and words and word placement. In our analysis we analysed speech rate, repetition, word placement and word duration. The latter four items based on an analysis of transcripts and recordings as well as considering the number of nouns extracted using the heuristics discussed in section 4.8 above. In order to assess learning performance of the robot we studied how well it classified words and how well items in the sensorimotor stream separated to focus meaning on the correct attribute. These are discussed below.

5.1 Speech Rate

Considering first the speech rate, this was calculated in words per minute (WPM) based on the number of words used by each participant in each session. The WPM varied on average between 80 and 132. However, 5 of the participants slowed their speech during the set of interactions, 4 ending with a lower WPM in the final session than in the first session. The other 3 participants varied their speech rates during the sessions but ended with a higher WPM by the final session (see figure 6).

Although the small participant sample size does not allow any firm conclusions to be drawn, indications of gender and child caring experience may be relevant for future larger scale studies. The 5 participants who slowed their speech rate were all female, 4 with direct child caring experience whereas the other participants were all male with no direct child caring experience. This may indicate that teaching a robot *as if* it were a young child may be more amenable for those already predisposed to such undertakings with a human child.

5.2 Repetition

An objective measure of repetition is to assess how many new words (of all types including nouns) are added between interaction sessions. If the number of new words decreases then it is likely the same words are being used between sessions. Similarly an analysis of words dropped (used in one session but not the next) would indicate, if increasing, that words are not being repeated, if decreasing, then the same words are being reused. Figure 7 shows the average number of new and dropped words between the sessions and indicates that repetition is being used (and increases) during the sessions. Note however that there is a sharp increase in the number of

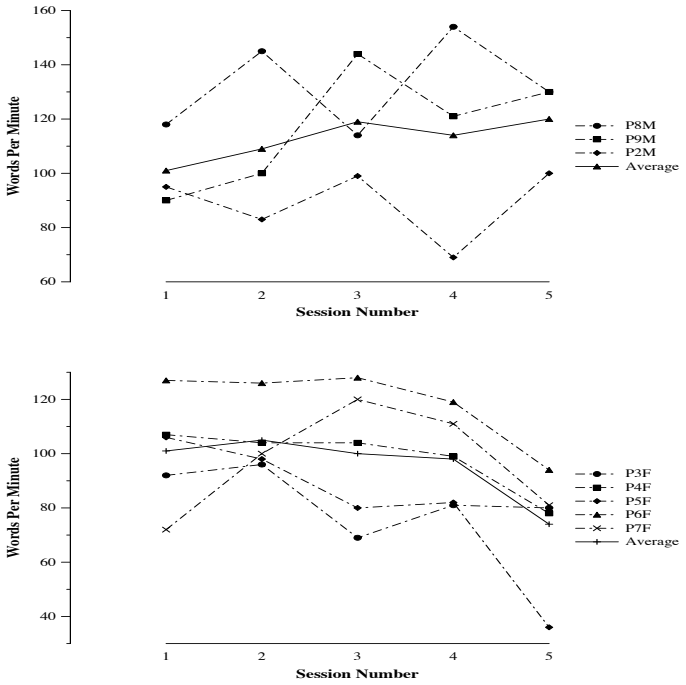

Figure 6. ***Speech rate.*** *The graphs shows the speech rate for participants over the five session. Top graph (all male) shows words per minute (wpm) varying with a general small increasing trend. Bottom graph (all female) shows wpm general trend decreasing. Trends line based on average by session shown as solid line.*

words dropped between the first and second sessions. This may indicate that once the robot starts to speak the human partner realizes the limited extent of its understanding and makes adjustments, in this case reducing the size of his/her expressed vocabulary.

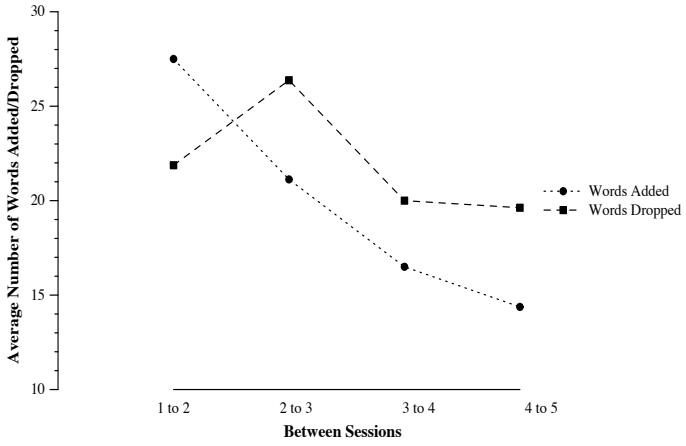

Figure 7. ***Words Added/Dropped Between Sessions.*** *The average numbers of words added and dropped between sessions for all participants. The number of new words decreases between all of the sessions. The number of words dropped increases sharply between the first and second sessions and then starts to decrease (see text). .*

5.3 Word Duration and Placement

To assess whether participants actually placed the salient words referring to shape description at the end of their utterances and with a longer than average duration we considered the performance of the heuristic described in section 4.8 above. We compared the words extracted by the heuristic against all possible nouns used by each participant in each session. If the participants regularly placed the shape name at the end of an utterance (and with higher than average duration) the extraction rate should be higher than average. Figure 8 shows the average for this measure for each participant. The results indicate that for all participants this effect occurred with the majority placing nouns at end of utterance and with longer than average duration over 80% of the time.

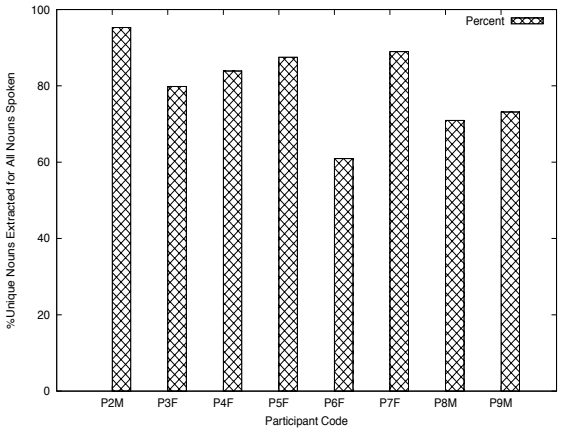

Figure 8. ***Noun Extraction.*** *The graph shows for each participant the performance of the heuristic of extracting words based on duration and end of utterance placement.*

5.4 Classification Performance

To assess the robot's learning ability each of the memory models was analysed using leave-out-out cross validation [25, pages 151-152] to predict classification performance (see figure 9). This method was used as it gave an unbiased and objective measure of classification accuracy over all of the participants sessions.

For all participants, performance according to the classification ability acquired by the robot had improved by the fourth session but started to worsen by the fifth session. This decrease in classification performance is due to the changes in the interaction style of the human. By session 4 the participants had noticed that the robot had started to correctly and consistently recognize the shapes, thus by session 5 their emphasis changed such that rather than place nouns at the end of the sentence various forms of reward utterances were used. For example, rather than continue saying 'this is a square' when the robot correctly and consistently recognized the square (and said 'square') the human would typically say 'well done' or 'good boy' or 'clever Kaspar'. This has the effect of diluting the previous association between the shape and the 'correct' word and thus reduced the classification performance. In this experiment we were not attending to such reinforcement signals, however the effect emphasized the fact that the participants were altering their behaviour in line with their perception of shared understanding between themselves and the robot.

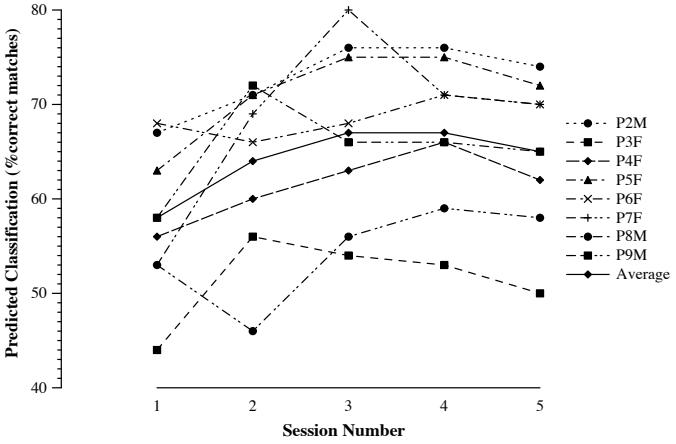

Figure 9. ***Predicted Classification Performance.*** *Classification performance by participant generally increases between the first and the fourth sessions. A decrease in performance between sessions 4 and 5 indicates that the human focus of the interaction has changed (see text).*

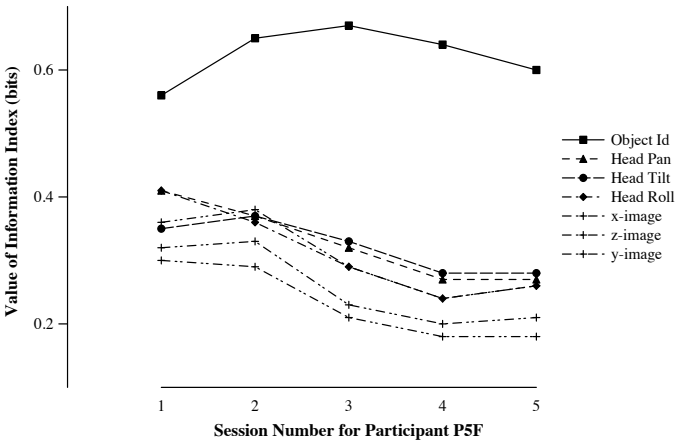

Figure 10. ***Separation of Meaningful Variables.*** *The graph shows how during the sessions the sensorimotor attributes which contributes most to the meaning of the word separate from other less meaningful attributes. This is an example from participant P5F.*

For the majority of the participants, there was also registered an increase in classification performance between the first and second sessions. This may be because the first feedback signals are being received from the robot in the second session (i.e. in the first session the robot is silent, in subsequent sessions it starts to speak). The robot speech feedback effectively signalling its level of understanding to the human tutor who subsequently adjusts his/her interaction appropriately. An alternative explanation is that participants had simply become familiar with the experiment, however participants P8M and P6F were the exceptions with a large performance drop of 13% for P8M and a marginal 2% drop for P6F, respectively. We believe that the large decrease exhibited by participant P8M was due to his interaction style i.e. not being engaged in the interaction. This behaviour becomes more apparent in the analysis below.

5.5 Identifying Meaningful Sensorimotor Attributes

Considering the fourth research question above (section 3). Here we are assessing whether sufficient information is available through the interaction to correctly enhance the key sensorimotor attributes (here the 'object id' returned from the ARToolKit system) and thus weight for likely selection via the *k*NN algorithm. This is effectively indicating that a particular set of attributes are more closely associated with the given word and, in our interpretation, more 'meaningful' to the robot. In figure 10 we show an example of this effect in detail for one participant (P5F), however the separation occurred for the majority of the participants in the study with varying degrees of success.

Figure 11 shows a summary for all participants. In this graph the values for 'object id' and the head and image proprioception sensorimotor attributes have been averaged over the 5 sessions. Note that for all but 2 participants (P4F and P8M) the meaningful variable (object id) separates from the other less meaningful variables. Participant P8M's interaction with the robot was characterized by both the highest speech rate (a WPM value of 132 vs an average of 102), the highest number of unique words used per session (90 vs an average of 60 for the other participants) and the highest number of nouns used per session (22 vs. an average of 13). Our interpretation of these figures suggests that the participant had not changed his interaction style to accommodate the robot's understanding. This is also reflected by the drop in classification performance over the initial sessions noted above (see 5.4).

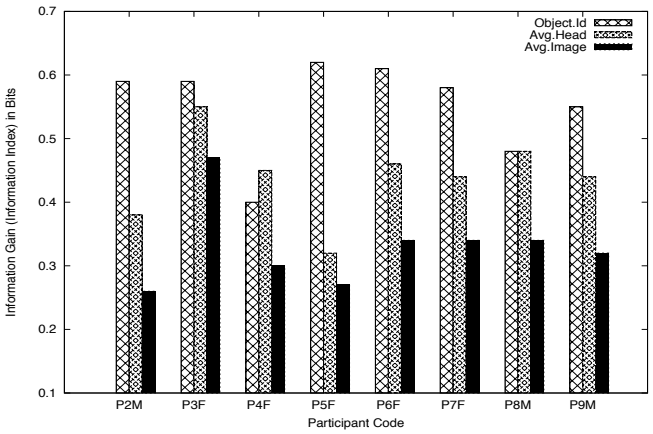

Figure 11. ***Separation of Meaningful Variables By Participant.*** *For each participant the separation of the average value of the object sensorimotor attribute vs. average head and image proprioception attributes. The key sensorimotor attribute weighting the classification of words via the* k*NN algorithm is the object id. The object id weighting separates from the other sensorimotor attributes as the sessions progress.*

Participant P4F achieved a classification performance similar to the other participants (average of 0.61 vs 0.64 for other participants) however the separation of the key object id attribute differed considerably. This would indicate that although the participant had engaged with the robot effectively she had not drawn sufficient attention to the object (thus she spoke about the object when the robot was attending elsewhere). The classification therefore indicated that the robot had made an association between head proprioception and the phoneme/word string rather than object and phoneme/word string. This was confirmed by subsequent review of the video for this par-

ticipant and showed Kaspar2 'saying' words regardless of where it was attending.

5.6 Summary

Returning to the original research questions, we considered whether firstly the human interaction partner would engage in a verbal style similar to that of Child Directed Speech. Indications are that the participants generally talked more slowly and were more repetitive emphasizing the objects via word placement and word duration. Secondly, we asked whether their robot-directed speech would change as the linguistic capability of the robot improved. There did indeed appear to be a definite change once it was accepted by the human partner that shared understanding between themselves and the robot had been achieved. Thirdly, in considering the robot's capacity to learn it was clear that the robot's ability to recognize and correctly classify shapes improved during the initial 4 sessions but decreased in the final session. We believe that this final decrease was due to the human perceiving shared understanding and thus subsequently deemphasizing previously salient words. The salience detection heuristic was not designed to cope with this and thus the subsequent classification performance decreased. Finally, the robot was able to attach sensorimotor meaning to the presented object and this was further highlighted in the robot's ability to distinguish which feature was important in identifying the presented shape from the set of attributes in the sensorimotor stream.

6 CONCLUSION

This study has attempted to demonstrate how interaction between a human and a robot can allow the robot to attach meaning to lexical items used by the tutor referring to a series of shapes. We have shown how the robot can derive meanings using some simple heuristics, developed with reference to human child language acquisition. Importantly no restrictions are placed on the human interactor other than to describe the shapes *as if* the robot were a young child. What is important is that the human adopts a language style similar to that employed in child directed speech and that some form of rudimentary shared reference is employed. This being the case the robot will learn the semantics of the shape words, based on its own sensorimotor feedback, with relatively few presentations.

Clearly, the case study discussed above is limited in that the number of participants is statistically relatively small (8 in total). We are therefore cautious in suggesting that the results are in any way conclusive. However we do believe that the results and the computational mechanisms outlined indicate interesting directions and novel approaches for future studies.

ACKNOWLEDGEMENTS

The work described in this paper was conducted within the EU Integrated Project ITalk ("Integration and Transfer of Action and Language in Robots") funded by the European Commission under contract number FP7-214668.

REFERENCES

[1] ARToolkit. http://www.hitl.washington.edu/artoolkit, 2003. [last visited on 30 June 2008].

[2] R N Aslin, J Z Woodward, N P LaMendola, and T G Bever, 'Models of word segmentation in fluent maternal speech to infants', in *Signal to Syntax*, eds., J Morgan and K Demuth, Lawrence Erlbaum, (1996).

[3] Paul Bloom, *How Children Learn the Meaning of Words*, MIT Press, 2002.

[4] Noam Chomsky, *Aspects of the theory of syntax*, MIT Press, Cambridge,MA, 1965.

[5] Eve V. Clark, *First Language Acquisition*, Cambridge University Press, Cambridge,UK, 2nd edn., 2009.

[6] Walter Daelemans and Antal van den Bosch, *Memory-Based Language Processing*, Cambridge University Press, 2005.

[7] Kerstin Dautenhahn, Chrystopher L. Nehaniv, Michael L. Walters, Ben Robins, Hatice Kose-Bagci, N. Assif Mirza, and Michael Blow, 'Kaspar - a minimally expressive humanoid robot for human-robot interaction research', *Applied Bionics and Biomechanics, Special Issue on 'Humanoid Robots'*, **6**(3), 369–397, (2009).

[8] eSpeak. http://espeak.sourceforge.net/, 2007. [last visited 31 July 2009].

[9] S. Harnad, 'The symbol grounding problem', *Physica D.*, **42**, 335–346, (1990).

[10] Carolyn B. Mervis and Laurel M. Long, 'Words refer to whole objects: Young children's interpretation of the referent of a novel word.', in *Paper Presented at Biennial meeting of the Society of Research in Child Development, Baltimore, MD*, (1987).

[11] MicroSoft. Microsoft speech technologies. http://www.microsoft.com/speech/speech2007/default.mspx, 2007. [last visited 30 June 2009].

[12] Chrystopher L. Nehaniv, 'Meaning for observers and agents', in *IEEE International Symposium on Intelligent Control/Intelligent Systems and Semiotics (ISIC/ISAS'99)*, pp. 435–440, (1999).

[13] Chrystopher L. Nehaniv, Kerstin Dautenhahn, Jens Kubacki, Martin Haegele, Christopher Parlitz, and Rachid Alami, 'A methodological approach relating the classification of gesture to identification of human intent in the context of human-robot interaction', in *14th IEEE International Workshop on Robot and Human Interactive Communication (Ro-Man 2005)*, pp. 371–377, (2005).

[14] openCV. http://opencvlibrary.sourceforge.net, 2006. [last visited 30 June 2008].

[15] N. Otero, C. L. Nehaniv, D. S. Syrdal, and K. Dautenhahn, 'Naturally occurring gestures in a human-robot interaction teaching scenario', *Interaction Studies: Social Behaviour and Communication in Biological and Artificial Systems*, **9**(3), 519–550, (2008).

[16] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.

[17] Deb Roy, 'Grounding words in perception and action: Computational insights', *Trends in Cognitive Sciences*, **9**(8), 389–396, (Aug 2005).

[18] Deb Roy and Alex Pentland, 'Learning words from sights and sounds: A computational model', *Cognitive Science*, **26**, 113–146, (2002).

[19] Jacqueline Sachs, Barbara Bard, and Marie L. Johnson, 'Language learning with restricted input: Case studies of two hearing children of deaf parents', *Applied Psycholinquistics*, **1**, 34–54, (1981).

[20] J. Saunders, C. L. Nehaniv, K. Dautenhahn, and A. Alissandrakis, 'Self-imitation and environmental scaffolding for robot teaching', *International Journal of Advanced Robotic Systems*, **4**(1), 109–124, (March 2007). special issue supplement on Human-Robot Interaction.

[21] Luc Steels, 'The origins of syntax in visually grounded robotic agents', *Artificial Intelligence*, **103**(1-2), 133–156, (1998).

[22] SysMedia. Sysmedia word and phoneme alignment software. [Last visited 31 July 2009], 2009. http://www.sysmedia.com.

[23] TimBL. http://ilk.uvt.nl/mblp/, 2005. [last visited 30 June 2009].

[24] Michael Tomasello, *Constructing a Language: A Usage-Based Theory of Language Acquisition*, Harvard University Press, 2003.

[25] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*, Elsevier, Morgan Kaufmann, San Francisco, 2005.

[26] Ludwig Wittgenstein, *Philosophical Investigations (Philosophische Untersuchungen)* – German with English translation by G.E.M. Anscombe, Basil Blackwell, 3rd edn., 1968. (first published 1953).

[27] C. Yu and D. Ballard, 'A multimodal learning interface for grounding spoken language in sensorimotor experience', *ACM Transactions Applied Perception*, **1**, 57–80, (2004).