Robot Acquisition of Lexical Meaning - Moving Towards the Two-word Stage

Joe Saunders, Hagen Lehmann, Frank Förster and Chrystopher L. Nehaniv
Adaptive Systems Research Group, School of Computer Science
University of Hertfordshire, Hatfield, UK, AL10 9AB
Email: j.1.saunders@herts.ac.uk

***Abstract*—We report on experiments and analyses dealing with the acquisition of lexical meaning in which prosodic analysis and extraction of salient words are associated with a robots sensorimotor perceptions in an attempt to ground these words in the robots own embodied sensorimotor experience. We focus here on three key areas, the selection of salient words based on prosodic clues, expression of words by the robot at a two-word stage to reflect learning and grammatically correct presentation, and an in-depth analysis of the relationship between words and the robots sensorimotor perceptions.**

I. Introduction

In these experiments and analytical studies we extend the work described by Saunders et al. [1], [2] and take inspiration from earlier studies by Roy [3]. We base the research on the principle that language acquisition in humans is fundamentally linked to social interaction and that children acquire language through interaction with others, typically their mothers, fathers, siblings or carers. Central to this idea is that processes in social interaction in early language acquisition provide sufficient bias for learning to occur and that meaning is derived from the association of salient words with embodied sensorimotor feedback as used in the context of embodied interaction. In the work described here we assume that segmentation of the speech of the carer into words is possible (and taken as a given), this then allows us to highlight salient portions (words or strings of phoneme or syllables) in the interaction. We use the techniques described below and deploy them in a humanoid robot by combining shared reference, word salience and embodied sensorimotor processes to provide the robot with the means to derive lexical meaning.

II. Child Directed Speech

In order to 'set the scene' we briefly summarise some of the key ideas from child language development that have inspired us in these studies.

Child Directed Speech (CDS) is a form of speech specifically tailored by adults to the perceived level of linguistic skill of the child [4]. In the case of acquisition of English as a first language, the following properties of linguistic interactions with children have been reported in the literature and inform the work presented here.

Adults typically speak more slowly, often reformulate their own and their childrens utterances and prominently draw attention to salient words via changes in pitch, pause duration, word placement and word duration. Pause duration has been measured in CDS [5] and together with falling prosodic intonation has been shown to be a reliable end of utterance marker for infants (both of these factors are used in the studies below to mark end of utterance boundaries).

Adult carers typically talk to infants more slowly and emphasize words by lengthening vowels [6] and this stretching out of words together with raised pitch is used to attract infant attention. Adults also use limited constructions and repetition to introduce new words with the new word often in the final utterance position [7], [5], [8].

Infants are exposed to a high number of ambient speech events, being exposed to as many as 7,000 utterances per day on the average [9], however they actually learn new words with very little exposure. Given that object/word learning is achieved by infants in very few episodes and that simple statistical association is not probable (as the co-occurrence of words and objects does not naturally occur frequently enough) the infants learning experiences must be biased in some other way. This is thought to occur via shared intentional reference. The utterances of the adult together with reinforcement via forms of affective feedback (prosodic features of CDS) and communicative success/failure of shared intentional reference for the infant whilst situated in context, together allow enough bias for fast learning to take place. Also the ability to form perceptual and conceptual categories of 'similar' objects and events and the ability to form sensory-motor schemas from recurrent patterns of perception and action potentially bias the learning process.

In the experiments described below we exploit these ideas, specifically pitch, energy (effectively volume), speech rate, pause duration and word duration to help our robot learner to assign lexical meaning to a series of words occurring in utterances about coloured shapes and differently sized boxes presented and talked about by a human tutor using unrestricted speech. Language acquisition as a social process can also be considered in terms of language games [10]. Following Wittgenstein [11], any derivation of meaning must ultimately be evidenced by appropriate embodied action in such language games (cf: [12]).

III. Studies in Robot Learning of Lexical Semantics

We report on the results of human-robot interacation experiments which attempted to progress from one-word learning (described in [1]) to two-word learning. As part of this experiment factors such as the effectiveness of the method of salient word extraction within utterances was tested and a further study made on which sensorimotor features the robot actually associated with salient words, whether these features can give clues to the appropriate temporal ordering of words (in English) and a detailed analysis of the speech of the human participants in order to give indications of what made some interactions more successful (based on the success of the robot word learning).

A. Experiment Outline

In these experiments we use a similar scenario as previously described in [2] whereby participants are asked to teach the iCub robot (called DeeChee), using their own, unrestricted speech, about a series of shapes drawn on three boxes. Each box is a different size, and each have six shapes (moon, star, arrow, cross, heart and circle). The shapes are coloured red, blue or green. The participant is told to treat DeeChee as a young child and to specifically teach DeeChee about the shapes, colours and sizes of the boxes (no mention of the shape names, colours or size adjectives are given to the participant). An experiment in progress is shown in fig. 1.

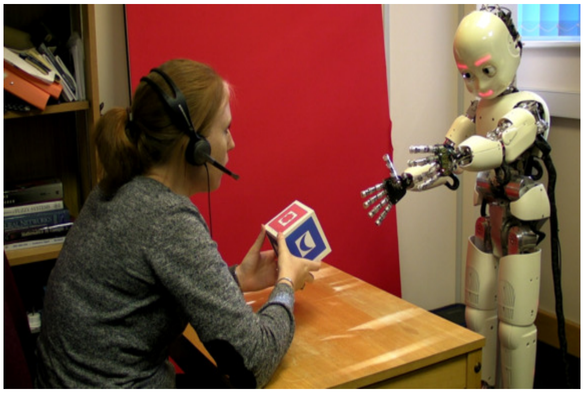

Fig. 1. Each participant was asked over five 2-minute sessions with the robot to explain the shapes, sizes and colours of the boxes as if the robot were a small child. From session 2 onwards the robot responds to the presented shapes by querying its association between current sensorimotor experience and salient words spoken by the participant in the previous sessions comprising its sensorimotor experiential history, and then expresses any chosen salient words according to its history of interaction with the particular participant.

If the robot detects a shape, it 'smiles' and then attempts to track it either making a deictic gesture (pointing), or simply moving its arm and hand towards the object. Either or both arms can move dependent on the location of the object in front of the robot. When no object is detected, the robot lowers its arms, looks 'sad' and focuses towards the participant making small random head and eye movements. This effectively engages the robot with the human participant. Each participant has five 2-minute sessions with the robot typically spread over a 2 week period. In the first session the robot does not 'speak'. However, from the second session onwards the robot may utter none, one or two words on each presentation of a box. The utterances of the robot are dependent on what it experienced in the previous session(s) both in terms of its perceptions and the salient words used by the participant during those sessions. Further explanations of the robots behavioural and learning algorithms are detailed in [1], [2], however explanations will be briefly restated where appropriate in the sections below.

The study was carried out in spring/summer 2011 and comprised 9 participants (6 female, 3 male) with ages ranging from 25 to 63. Typically the participants were either PhD students or university administrative staff. None were connected with robotics research. In total 45 two-minute interaction sessions were carried out.

IV. Research Questions

In an attempt to progress from one-word learning to rudimentary two-word learning a number of factors have to be considered. Firstly, in any given utterance expressed by a participant, which words are salient, and will these salient words provide the robot with an appropriate bias to ground the words in its own sensorimotor perceptions? In this study we denote these as 'meaningful' words. Here 'meaningful' words are those which could be considered to be associated with the robot's sensorimotor dimensions e.g. the word 'red' would be a meaningful word because it could be associated with the robots vision colour dimensions, however, for example, the word 'clever' would not be meaningful as the robot at present does not have any facility for sensing cleverness. We expressed this as our first research question:

1. Is it possible to extract 'meaningful' words from the utterances expressed by the participant and how effective is this extraction?

In this context the 'effectiveness' metric is defined as the number of words used by the participant which could possibly be associated with a sensorimotor dimension, e.g. the words 'small' and 'little' could both be used to describe size and be associated with the numeric value denoting smallness. Given that the robot has 3 dimensions for shape, colour and size and that there are 6 shapes, 3 colours and 3 sizes, we measured 'effectiveness' as the success in extracting words used within any session which described these items.

For our second research question we focused on word order. We expected that the word order of two word utterances should reflect the conventions used in the target language. In these studies we have focused our prosodic extraction algorithms specifically to English speakers and as such we would expect that any two words expressed by the robot would follow the normal English word order e.g. 'blue moon' rather than 'moon blue' [1].

2. Given the received temporal ordering of salient words in utterances expressed by the participant, would this ordering

[1] However, the latter ordering could occur in contexts of predication in English as well, rather than within a noun phrase with where adjectives precede the noun head.

give clues as to the usual word order in English?

If this were the case, we would expect that in any two word order (and given our limited scenario), adjectives would precede object names and that the robots sensorimotor perceptions would favour dimensions associated with adjectives (e.g. colour, size) when expressing a first word but favour dimensions associated with nouns (e.g. shape) for the second word when re-experiencing the object. If this dimensional weighting were to occur (and given that we would not ascribe any pre-categorization to these dimensions i.e. we would not label the colour dimension as an adjective) this might provide a possible mechanism for eliciting learning of the typical word order in a language.

For our third research question we noted in our original studies described in [1], [2] that there was variation in the participants relative success in teaching the robot shape names. We suggested that this was partly due to lack of engagement with the robot (e.g. not emphasizing salient words, as would be the case for child-directed speech). In this study we have used more complex shape stimuli (with size and colours) and wanted to more fully study how effective the participants were in teaching the robot. Therefore our third question was:

3. What are the factors which make a participant more successful in teaching a robot about shapes, colours and sizes of boxes?

We needed to also distinguish between technical factors affecting the robots ability to learn (e.g. not recognising or mis-recognizing the shapes due to failure in the vision system, associating items outside the temporal window of the utterance where shapes are spoken about but a different shape is presented to the robot) and what we call 'human' factors e.g. not engaging with the robot or simply talking too fast, too slowly or simply not talking about the 'right' things.

V. Software Architecture

We will briefly describe the architecture and key enhancements made for this study but for details of the underlying architecture please refer to [1], [2]. In carrying out these studies we used an existing social learning architecture called ROSSUM [13]. The architecture has been previously used to allow a robot to learn scaffolded behaviours via directed learning from a human. This experiment used the iCub robot (see fig. 1) and employed the inverse kinematics library (Pattacini, 2010) available for the iCub to control the head, eyes and arms of the robot.

A. Shape Recognition

Shape recognition was via an enhanced version of the ARToolKit system [14] in which colour and size were also made available to the robot. The objects were pre-learnt using ARToolKit, thus the iCub could detect these objects and recognize that they were individual entities in the world (shape classes comprised of a star, arrow, circle, cross, moon and heart which corresponded to the integers 1-6 in a sensory stream for object class), however no other meaning was attached to them. A similar coding scheme was used for colour sensor values (integers 1-3 corresponding to red, blue and green) and for size values (integers 1-3 corresponding to large, medium and small). 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, it reflects a 'whole object' bias found both in children and adults (see([9], Chapter 4) and [15].

B. Salient Word Recognition

Following each session the participants speech was manually transcribed and then automatically aligned against the speech signal to yield a set of timed phonemes with word markers. This was achieved using a combination of software components from SysMedia [16] which carried out the initial word alignment and the University College London SFS system [17] which converted the timed aligned words to phonemes and realigned them. In future research we will be investigating more direct methods of alignment and extraction (see for example [18]). Following the alignment process the timed phoneme/word file was processed by Huang, Chen and Harpers Prosodic Feature Extraction Tool (PFET) [19]. This uses Praat [20] as its underlying analytical engine and allowed the extraction of various prosodic statistics at the word level via analysis of the phonemic input. Note that the PFET system exploits the Praat systems default pitch stylization, voicing and octave jump settings. Clearly such a word-based approach is relatively simple, however we exploit the fact that English is a stress-timed language and that our robot directed speech data did not contain many multi-syllable words.

We identified individual phrases based on pause duration and word duration. Average pause duration is computed as the sum of each pause between words divided by the total number of words. If the pause between words exceeds the average pause duration, then an end of phrase boundary marker is inserted after that word. Note that even if the speaker hesitates, the phrase is still considered to be complete up to the hesitation. The phrase may not be grammatically complete, however the prosodic information is still considered to be present. A further segmentation of each subsequent utterance was then carried out. This is based on studies indicating that (at least in English) persons using CDS favour introducing new information at ends of utterances [21], [4]. A signal for the new information is extended word duration. We placed a phrase boundary marker after a word whose duration is larger than the first standard deviation of all words in that utterance. This had the effect of splitting longer utterances where pause duration was ineffective.

Using statistics provided by the PFET system above, we extracted values for maximum fundamental frequency ($f0$), duration and maximum energy occuring in each word. We identified salient words based on the pitch, energy and duration features of the word following the evidence for prosodic salience outlined in section II above. Within each utterance we normalized the values of $f0$, energy and duration and then multiply the normalized values together to give an overall measure of salience for each word. A word was then marked as salient if the normalized measure was larger than the average normalized measure based on all words in the utterance. This had the effect of highlighting words which have not only high pitch features but also extended duration or are said very loudly. For single word utterances we computed the first standard deviation of pitch, energy and duration respectively for the whole interaction session. If any of the pitch, energy or duration features were larger than the first standard deviation for the complete session for that feature then that word was selected as salient.

C. Attaching Meaning to Words

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 word/sensory-motor stream of the robot. The first step in this process was to merge the speech stream of the human (represented as a set of words with word boundaries) with the robots sensory-motor stream. This was achieved by matching the two modalities based on time.

This merger effectively associates what was said to the robot with everything else experienced by the robot at that time. Within this study the set of sensorimotor attributes was limited to the robots perception of object shape, size and colour (in previous studies we also used other physical and visual dimensions; however, in this study we wished to restrict the possible set of attributable meanings to an easily analysed set).

Note nevertheless that in this study we are dealing with a set of salient words extracted from each human utterance. The extracted words are in the order that they were expressed. For the purposes of allowing the robot to subsequently express two words we split the salient words into two tables. The first contained all salient words before the final salient word. The second table contained the final salient word. (Note that these are salient words and as such not every word in the utterance is selected and the final salient word may or may not be the final actual word in the utterance).

This resulted in two tables subsequently used for sensorimotor similarity matching (using k-Nearest Neighbour) in the robots execution phase. Having associated the relevant word with the utterance/sensorimotor stream, we then computed the information index [22] 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 dimension). This measure is used during the next interaction sessions (2-5) to weight the similarity measure of current vs. stored experiences. Note, in our experiments the robot learnt lexical semantics separately from each participant so that learning occurred in effect as if each participant 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).

The robot matched the resulting tables against its current sensorimotor perceptions and thus tried to find the most similar experience of when it 'heard' a word previously compared to what it was now experiencing. This has the effect of making the robot utter words which reflect both what it was taught (about objects, sizes and colours) and the order in which the words originally occurred. That is, by uttering any best matching words for each of the two tables, upon seeing a new colour/object shape (where it has seen both the colour and object before but not in this combination), the robot should express the correct attributes in a (proto-)grammatically correct order reflecting usage by the human it learned from.

VI. Results and Analyses

In this section we will present some results and general discussion of the research questions posed above. The first question posed was: *Is it possible to extract 'meaningful' words from the utterances expressed by the participant and how effective is this extraction?* The first issue to be addressed in answering this question is to consider what is meant by 'meaningful' words. To reiterate, here we consider meaningful words to be those words that could, in theory, be associated with a particular value within the robots set of sensorimotor dimensions. For example, the word 'moon' or 'crescent' are often used by participants in naming the moon shape. Both of these are meaningful words and could possibly be used by the robot to ground the 'moon' shape value in the shape dimension. However, words such as 'that', 'another', 'really', 'good' etc. have no direct meaning to the robot as it is not capable of attaching any one or combination of existing dimensional attributes to these words. Thus in order to answer the first research question required that a count of the possible set of meaningful words for each session and for each participant is made. This should then be compared with the words extracted by the salience algorithm (described in section V-B above) and the equivalent set of meaningful words counted. This gives a measure of the effectiveness of extracting meaningful words in the session. Additionally, a 'sanity' check of effectiveness can be made by comparing the set of meaningful words extracted against the set of words which would be randomly selected from the original session dialogue but in the same proportion as the algorithm selects. To put this more simply, we estimate how well the algorithm (based on pitch, energy and duration) does in extracting meaningful words and compare this to a random selection.

Each participant is identified by a unique code on the left of the table where F indicates female and M indicates male. Each participant‘s average over the five sessions is shown. For example, of all of the words expressed by participant F01, 27% were meaningful (using the criteria above). The salience algorithm extracted 46% of all words. This selection of 46%

TABLE I
Effectiveness in extracting meaningful words. Each row represents the average over 5 sessions of a given participant.
Note: ’Meaningful Words Extracted’ is shown as a % of ’Words Extracted’.

Participant	Meaningful Words	Words Extracted	Meaningful Words Extracted	Versus Random
F01	27%	46%	82%	46%
F02	20%	44%	79%	44%
F03	16%	47%	80%	47%
F04	29%	47%	83%	47%
M01	61%	52%	73%	52%
F05	15%	48%	75%	48%
M02	41%	49%	69%	49%
M03	32%	48%	80%	48%
F06	13%	46%	79%	46%
Average	28%	47%	78%	47%

of the words contained 82% of all the meaningful words, thus even though just under half of the words were selected, this portion contained 82% of the 27% meaningful words. If we selected 46% of the dialogue at random, this would yield 46% of the meaningful words. Therefore the algorithm exceeds random selection by 36%.

This measure of effectiveness of automatically extracting meaningful words indicates firstly, that the algorithm is very successful at extracting salient, meaningful words. But secondly, it also shows that the participants in most cases are prosodically highlighting such words in a manner similar to that used in child-directed speech. However, individual variations are interesting to note, for example, for participant M01 around 61% of words used were meaningful, but the algorithm was rather less successful at extracting such words. This indicates (and is verified by reviewing the dialogue transcripts), that this participant was rather ‘unnatural’ in his interaction with the robot, simply saying ‘red star, blue arrow’ rather than more extended utterances for example ‘here is a lovely red star, like you see in the sky’. The algorithm was less successful because there was less prosodic emphasis, the words being uttered in a rather neutral way.

Our second research question was: *Given the received temporal ordering of salient words in utterances expressed by the participant, would this ordering give clues as to the usual word order in English?* There are different ways of expressing simple sentences. For example, ‘This is a red star’ and ‘This star is red’. Typically in English the adjective precedes the noun (as in the first sentence) when drawing attention to objects with multiple properties. Thus the first sentence is typically implying that the object is a ‘red star’ whereas the second sentence implies that the listener knows about ‘stars’ but the speaker is saying that this particular star is ‘red’. Grammatically, in the first sentence ‘red’ is a modifer that can be thought of as incidently selecting the referent, whereas, in the second, ‘red’ is a property being predicated of the (known) object shown to the robot. Therefore in the interactions with the robot in this study, where the robot might be assumed to know very little about the shapes, colours and sizes, we expected that there would be more use of the sentences of the first type above. However, the robots learning mechanism includes no such grammatical presuppositions in its design, so that learning proceeds as described above based on usage by the participants in interactions over several sessions. Since this processs is iterative, that the robots utterances influence what the participant says from the second session onwards potentially creating regularities and habits, and thus ‘grammar’ in the sense of Wittgenstein [11] in the resulting language games are particular to the embodied setting of the particular participants interaction history with the robot learner.

In the study we split all pre-final salient words into one table and the final salient words into a second table, with both tables containing all of the sensorimotor dimensions and the associated word for each row derived from our temporal matching of the robots sensorimotor information and the participants salient utterances. Both of the tables were subject to an information index (mutual information) analysis which gave a numeric weighting to each dimension in each row of each of the tables. The weightings are normally used in the execution phase to more heavily emphasize one dimension over another. If the supposition that modification would occur more frequently than predication in salient words of the participant‘s speech directed to the robot, one would thus have expected that the first table would normally weight the dimensions for colour and size, whereas the second table would more heavily weight the shape dimension. This would cause the robot, when, for example, seeing a ‘red star’, to match its sensor values for ‘red’ and ‘star’ against the first table and then against the second. As the ‘colour’ dimension may be more heavily weighted in the first table, the ‘red’ sensor value would dominate and thus yield the word most associated with that dimension (hopefully the word ‘red’). Similarly, the shape dimension would be more heavily weighted in the second dimension, and hopefully yield ‘star’.

If this were the case then we would expect that over time (and during the 5 sessions), the dimension appropriate to the ‘usual’ constructional use of the dimensional items (i.e. colour, size and shape) should occur, and thus colour would for example typically precede shape (and thus be more heavily weighted in the first table than the second).

The graphs in fig. 2 show the difference between the total information index for all 9 participants over the five sessions. The blue line is the colour dimension (upper graph) or the size dimension (lower graph) and the red line is the shape dimension. The result shows how the colour and size dimensions are more heavily weighted in the first table and the shape dimension more heavily weighted in the second table (indicated by a widening difference between the two). The implication being that by studying how the robots sensorimotor dimensions alter over the course of the interactions we may be able to derive the typical usages and ordering of those dimensions within the target language, which in turn would reflect its grammar. Note however that the effect is less

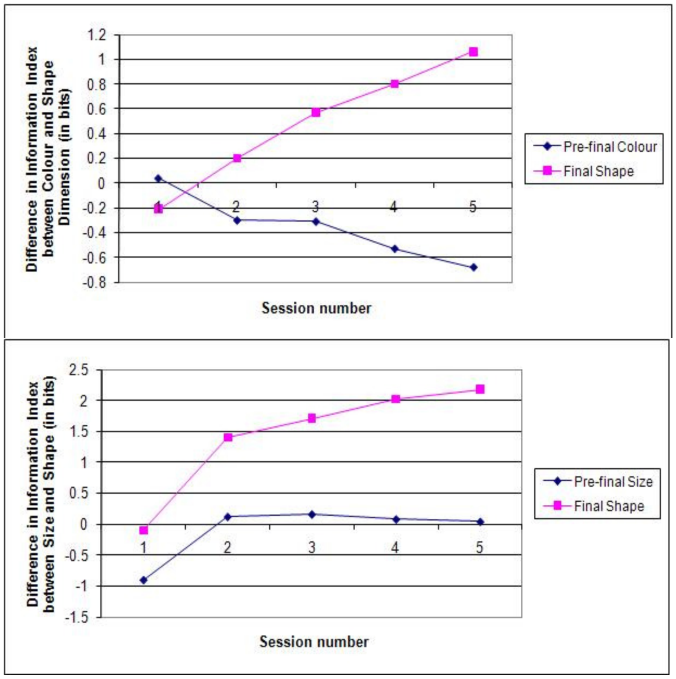

Fig. 2. The graphs are a summary of differences in the participants 'colour' and 'shape' (top) and 'size' and 'shape' (bottom) dimensional information index weighting at each session. The red line shows the difference between the dimensions of colour or size vs. shape in the first sensory/salient word table . The blue line shows the difference between the dimensions of colour or size vs. shape in the second sensory/salient word table. Each point is calculated by summing the information index (mutual information) for the relevant sensory dimension (e.g. colour, size, shape) for that session over all participants. Information index looks at each dimension in isolation and then measures how much information that dimension contributes to our knowledge of the correct class label (words in this case). See [23], page 23 for more information on information index. Here the information index value for each participant is totalled at each session for each dimension and the differences displayed.

distinct for the 'size' dimension (lower graph). We believe that this was because participants spoke about size far less often than colour and may have been due to the fact that the boxes differed in size, but the shapes did not. Therefore it was difficult when discussing shapes to associate this with the box size e.g. when talking about shapes the participant was not able to say 'this is a big blue star', but rather say 'this is a blue star on the big box'. Thus shapes and sizes were not easily associated.

Our third research question asked: *What are the factors which make a participant more successful in teaching a robot about shapes, colours and sizes of boxes?* In this question we need first to define what is meant by success in teaching the robot. Our definition of success here is to analyse how many words out of the set of possible meaningful words (see research question 1) the robot was able to correctly associate with its appropriate sensorimotor dimensions. For example, if the unique set of distinct meaningful words used by the participant was 16, and the robot correctly assigned the correct sensorimotor dimension and values to 12 words, then success would be 12/16 = 75%.

In terms of the 'factors' contributing to success (or otherwise), we considered total number of words spoken, unique words spoken and the set of meaningful words expressed in both counts. Detailed results are shown in table II.

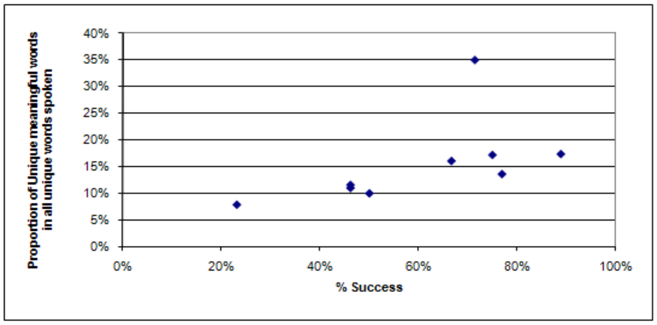

Fig. 3. The points on the graph show the value of each participants success in making the robot learn the meaning of words versus the number of unique words spoken as a proportion of all unique words spoken.

Summary results are shown on the graph in fig. 3. Table II indicates that there are some variations in the participants behaviour towards the robot. For example, participant F06 used many words; however, few of them were meaningful and the robot was much less successful at making correct associations when learning from this participant. Also participant M01 used many meaningful words but was not the most successful. A main indicator of success appears to be the number of unique meaningful words as a proportion of total unique words used in the sessions. This can be thought of as the *density* of meaningful words rather than the actual number of words, or the actual number of meaningful words. Thus (apart from the outlier of participant M01 discussed in the first research question above) the trend appears to suggest that factors which make a participant more successful in teaching a robot about shapes, colours and sizes of boxes is the density of meaning. To some extent this seems an intuitive result in suggesting that the robot will learn more if we use the words denoting shapes, colours and sizes more than any other words spoken. But two factors are in operation, firstly the meaningful word density plus the ability to extract words that are salient. Thus the participant has to prosodically emphasize the meaningful words as well as use them more densely. We should make clear however that these results are based on a small dataset of 9 participants over 45 sessions and are therefore only indicative at best, further research will be needed to confirm these results.

VII. Conclusion

We have presented an analysis of experiments which studied the acquisition of lexical meaning in a humanoid robot during interaction sessions with a human teaching it about shapes, colours and sizes using unrestricted speech. The human was asked to consider and treat the robot as an infant. In this study the number of participants were limited (9 in total), nevertheless over 45 individual human-robot interaction sessions were analysed. From these studies we have however shown that

TABLE II

Results from the 9 participants totalled over five sessions. Total words shows the total words used by the participant for the five session. Total Meaningful indicates the number of meaningful words used in all of the words spoken. Unique words and unique meaningful words are the number of distinct words of these types. The associated words column shows whether the meaningful words are correctly associated with the appropriate robot sensorimotor dimensions and values. Success is the percentage of correctly associated words from amongst the unique meaningful words.

	Total Words	Total Meaningful	%	Unique Words	Unique Meaningful	%	Associated	Success %
F01	627	170	27%	52	9	17%	8	89%
F02	1124	224	20%	96	13	14%	10	77%
F03	1045	166	16%	119	13	11%	6	46%
F04	1035	299	29%	113	13	12%	6	46%
M01	610	370	61%	40	14	35%	10	71%
F05	1044	155	15%	121	12	10%	6	50%
M02	540	219	41%	75	12	16%	8	67%
M03	811	259	32%	70	12	17%	9	75%
F06	1420	178	13%	167	13	8%	3	23%

it is possible to extract meaningful salient words from the humans speech pattern, for the robot to subsequently express these words in a grammatically correct two-word format when experiencing similar sensorimotor situations. Finally we have demonstrated that it is the density of salient meaningful words that allows effective learning to take place.

Our future studies will also consider the contingent aspects of the interaction [24], [25] in order to contrast the effectiveness of the acquisition of lexical meaning where a robot acts in a contingent or non-contingent manner, and further studies on the acquisition of linquistic negation [26].

Acknowledgment

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] J. Saunders, C. L. Nehaniv, and C. Lyon, "Robot learning of lexical semantics from sensorimotor interaction and the unrestricted speech of human tutors," in *Proc. Second International Symposium on New Frontiers in Human-Robot Interaction, AISB Convention, Leicester, UK*, 2010.

[2] J. Saunders, H. Lehmann, Y. Sato, and C. L. Nehaniv, "Towards using prosody to scaffold lexical meaning in robots," in *Proceedings of ICDL-EpiRob 2011*. IEEE, 2011.

[3] D. Roy, "Grounding words in perception and action: Computational insights," *Trends in Cognitive Sciences*, vol. 9, no. 8, pp. 389–396, Aug 2005.

[4] E. V. Clark, *First Language Acquisition*, 2nd ed. Cambridge,UK: Cambridge University Press, 2009.

[5] P. A. Broen, "The verbal environment of the language learning child," *Monograph of the American Speech and Hearing Association*, vol. 17, 1972.

[6] O. K. Garnica, "Some prosodic and paralinquistic features of speech to children," in *Talking to children: Language input and acquisition*, C. Snow and C.A.Ferguson, Eds. Cambridge, UK: Cambridge University Press, 1977.

[7] E. V. Clark and A. D. Wong, "Pragmatic directions about language use: Words and word meanings," *Language in Society*, vol. 31, pp. 181–212, 2002.

[8] C. A. Ferguson, D. B. Peizer, and T. E. Weeks, "Model-and-replica phonological grammar of a child's first words," *Linqua*, vol. 31, pp. 35–65, 1973.

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

[10] R. Brown, *Words and things*. New York: Free Press, 1958.

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

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

[13] J. Saunders, C. L. Nehaniv, K. Dautenhahn, and A. Alissandrakis, "Self-imitation and environmental scaffolding for robot teaching," *International Journal of Advanced Robotic Systems*, vol. 4, no. 1, pp. 109–124, March 2007, special issue supplement on Human-Robot Interaction. [Online]. Available: http://www.ars-journal.com/International-Journal-of-Advanced-Robotic-Systems/Volume-4/ISSN-1729-8806-4115.pdf

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

[15] C. B. Mervis and L. 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.

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

[17] M. Huckvale, University College London *et al.*, "Speech filing system," [Last visited 06 April 2011], 2011, http://www.phon.ucl.ac.uk/resource/sfs/.

[18] L. Pearl, S. Goldwater, and M. Steyvers, "Online learning mechanisms for bayesian models of word segmentation," *Research on Language and Computation*, vol. 8, pp. 107–132, 2011.

[19] Z. Huang, L. Chen, and M. Harper, "An open source prosodic feature extraction tool," in *Proceedings of the Language Resources and Evaluation Conference (LREC)*, 2006.

[20] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]." 2010, retrieved 29 March 2010 from http://www.praat.org/.

[21] A. Fernald and C. Mazzie, "Prosody and focus in speech to infants and adults," *Developmental Psychology*, vol. 27, pp. 209–221, 1991.

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

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

[24] K. S. Lohan, K. J. Rohlfing, K. Pitsch, J. Saunders, H. Lehmann, C. L. Nehaniv, K. Fischer, and B. Wrede, "Tutor spotter: Proposing a feature set and evaluating it in a robotic system." *I. J. Social Robotics*, vol. 4, no. 2, pp. 131–146, 2012.

[25] K. Lohan, K. Pitsch, K. Rohlfing, K. Fischer, J. Saunders, H. Lehmann, C. Nehaniv, and B. Wrede, "Contingency allows the robot to spot the tutor and to learn from interaction," in *Development and Learning (ICDL), 2011 IEEE International Conference on*, vol. 2, aug. 2011, pp. 1 –8.

[26] F. Förster, C. L. Nehaniv, and J. Saunders, "Robots that say 'no'," *European Conference on Artificial Life (ECAL 2009), Budapest, Hungary*, 2009.