A Constructivist Approach to Robot Language Learning via Simulated Babbling and Holophrase Extraction

Joe Saunders, Caroline Lyon, Frank Förster, Chrystopher L. Nehaniv and Kerstin Dautenhahn

Abstract— **It is thought that meaning may be grounded in early childhood language learning via the physical and social interaction of the infant with those around him or her, and that the capacity to use words, phrases and their meaning are acquired through shared referential 'inference' in pragmatic interactions. In order to create appropriate conditions for language learning by a humanoid robot, it would therefore be necessary to expose the robot to similar physical and social contexts. However in the early stages of language learning it is estimated that a 2-year-old child can be exposed to as many as 7,000 utterances per day in varied contextual situations. In this paper we report on the issues behind and the design of our currently ongoing and forthcoming experiments aimed to allow a robot to carry out language learning in a manner analogous to that in early child development and which effectively 'short cuts' holophrase learning. Two approaches are used: (1) simulated babbling through mechanisms which will yield basic word or holophrase structures and (2) a scenario for interaction between a human and the humanoid robot where shared 'intentional' referencing and the associations between physical, visual and speech modalities can be experienced by the robot. The output of these experiments, combined to yield word or holophrase structures grounded in the robot's own actions and modalities, would provide scaffolding for further proto-grammatical usage-based learning. This requires interaction with the physical and social environment involving human feedback to bootstrap developing linguistic competencies. These structures would then form the basis for further studies on language acquisition, including the emergence of negation and more complex grammar.**

I. Introduction

In learning to use language to communicate and manipulate the world around them, human children benefit from a positive feedback loop involving individual learning (by interacting with their hands and bodies with objects around them), social learning (via close interaction with parents and others), and gradual acquisition of linguistic competencies. This feedback cycle supports the scaffolding of increasingly complex skill learning and linguistic development giving the child ever greater mastery of its social and physical environment, as well as supporting the development of cognitive and conceptual capabilities that would seem impossible without language. Our work is aimed at realizing this same kind of feedback cycle supporting that scaffolding of behavioural, linguistic and conceptual competencies in robots. The purposes of doing this are not only to better understand possible mechanisms for such learning in humans, but also to achieve similar competencies in artificial agents and robots (even if they are not acquired by exactly the same routes). In this paper we report on our currently ongoing and proposed forthcoming experiments which employ the ideas above.

This work is inspired by the observed progress in language acquisition by human infants. Though we do not aim to simulate this development as a whole, we investigate certain mechanisms that could play a key role. We make the artificial assumption that we can isolate different developmental paths and examine them separately. Specifically, in our experimental scenarios, we initially model the acquisition of the phonetic form of words and holophrases without meaning, by employing simple learning mechanisms. Functional use and 'referential understanding' of utterances (e.g. registration of sensorimotor and environmental regularities), in a human-robot interaction context where joint-attentional framing and simple actions with objects are possible, will be introduced in subsequent steps as detailed below. Meanwhile, lexical/holophrastic learning continues and serves to bootstrap (1) learning of sensorimotor and interactional grounding of speech and behavioural skills, and (2) learning of the usage of learnt linguistic structures for the interacting robot to generate utterances that serve to manipulate its physical and social environment. The emergence of the capacity to use various forms of linguistic negation is also targeted in this experimental setting.

II. From Babbling to the Acquisition of Words and Phrases Without Meaning

Initially, the infant's ability to perceive and analyse acoustic signals is much greater than the ability to produce them. This contrasts with the mature speaker, who has matching perceptive and productive competencies, linked in mirror neuron processors. Cognitive capacity precedes productive abilities in speech. For instance, infants can use function words to help segment and analyse utterances before they produce them themselves [1, p. 201]. In an analogous way our Linguistically Enabled Synthetic Agent (LESA) will have an asymmetrical language competence. We aim for it to take in natural English and to respond with appropriate actions and spoken comments. However, its spoken output will be limited. The development of the ability to segment a speech stream into words and phrases overlaps with the acquisition of semantic understanding and the mastery of primary language structure, but here we initially just investigate the emergence of stable phonetic forms or strings independent of meaning. Our simulation aims to show how a

The authors are with the Adaptive Systems Research Group, Centre for Computer Science and Informatics Research, University of Hertfordshire, College Lane, Hatfield, Herts AL10 9AB, United Kingdom (email: {J.1.Saunders,C.M.Lyon,C.L.Nehaniv,K.Dautenhahn,F.Forster}@herts.ac.uk).

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.

978-1-4244-2763-5/09/$25.00 ©2009 IEEE

robot might learn to segment an utterance, an acoustic stream of sounds, that will be the basis for extracting meaningful elements [2, p. 39]. The starting point is taken as the stage at which infants typically produce canonical babbling, in the second half of their first year. This is the age at which they begin to show they are learning the phonetic characteristics of their own ambient language [1, p. 46], [3, p. 148]. Earlier the ability to distinguish different phonetic contrasts is universal, but sensitivity to some foreign phonemes is then lost or diminished – for instance the capacity of infants learning Japanese to distinguish between /r/ and /l/ as both sounds belong to the same Japanese phoneme [r] – while discrimination of native sounds improves [4].

Infants typically start to acquire vowel categories of their native language and combine them with certain consonants to produce characteristic repetitious strings of syllables *dadada*, *mamama*, *etc.* typical of babbling. This is the basis on which the infant will build its language production abilities [5, p. 50] as "[t]he first (pre-symbolic, pre-referential, context-limited) words produced reflect a match between the child's babbling patterns and adult patterns produced in a meaningful context" [6, p. 136].

In the acquisition of linguistic capabilities, e.g. mastery of word-usage frames and case markings, there are analogies with holistic utterances hypothesized as components of protolanguage in early evolutionary stages [7], [8].

From the start we will set up a dialogue, with turn taking, between a LESA and a human, or simulated human, teacher. We make certain assumptions and then see whether they suffice for aspects of autonomous development of language abilities in the LESA. The assumptions include:

- LESA has the intention to communicate.
- Communicative ability is learnt via interaction with a teacher.
- Perception and production of speech are based on simulated 'mirror neuron' type structures - i.e., the same elements reflect components of perceived speech and generate synthesized speech (cf. [9], [10]).

Recent research suggests that mirror neurons are like junctions, linking to various memory sites and potentiating action. Or, as Damasio says, they are "like puppet masters, pulling the strings of various memories" [11].

A. Memory structure

There is significant evidence that dual memory systems are needed for language processing, involving modular regions of activity as well as shared areas. On the one hand there is *explicit* "declarative" learning, in which there is joint attention between teacher and learner, and reference to objects, actions or relationships, sometimes described as item learning. On the other hand there is *implicit* learning of patterns and procedures, without intentional shared reference [12]. This dichotomy is also described as a ventral pathway specializing in object identification and whole word recognition in contrast to a dorsal pathway concerned with object interactions and phonetic decoding, sub-lexical processing [13], [14, p. 175-6]. An outline of stages in the development of linguistic competence up through holophrastic/lexical levels in each of these two modes is shown in Figures 1 and 2 (but realizing that there is probably a continuum between the modular and shared memory processes).

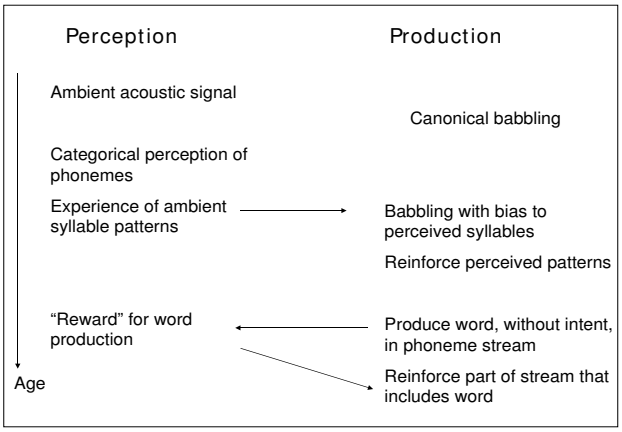

Fig. 1. ***Implicit, pattern learning***

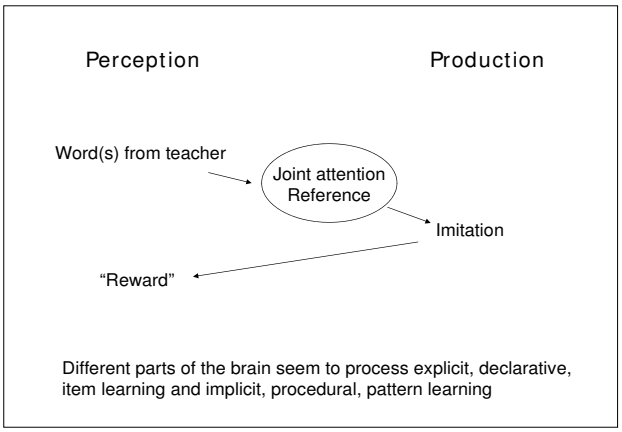

Fig. 2. ***Explicit, item learning***

Implicit learning comes first, and is then integrated with explicit, item learning.

B. Experiment I

1) Phonemics and Holophrasis Acquisition: For these preliminary investigations LESA will be a software agent. The teacher will be partly an actual human, partly a simulated human so that we can increase the quantity of input. From the start we aim to eventually have LESA 'understand' a co-operative but naive human – as evidenced by speaking and acting appropriately in context. Therefore, we do not just want to work with stylized sentences, but with the vagaries of natural language.

Using the SCRIBE corpus of phonetically rich English [15] a store of phonemes, diphones and demi-syllables is implemented, simulating a component of LESA's phonemic knowledge. Other sources could be used for different languages. The structure of the syllable as perceived by infants [1, p. 27] will be modelled. This is analogous to the

result of an infant constantly hearing English speech around him or her, not necessarily infant-directed speech (IDS). At this initial stage in our work the structure of memory and knowledge fusion processes is simplistic.

For the case of English, the LESA will start by combining certain consonants and vowels into syllables, using the range of phoneme combinations observed in infants acquiring English [16, p. 149],[17, p. 176], [18, p. 25], with appropriate weighted chances of being used. LESA will generate output strings of simulated babble in variable length utterances. The teacher will produce short English language utterances, typical of the length found in motherese as in the CHILDES database [19], which are converted and encoded into phoneme strings, where the phonemic alphabet includes a start/end-of-utterance indicator.

To begin with there is no correspondence between utterances: interaction between LESA and its teacher is confined to turn taking. Initially LESA's output is simulated babble, quasi-random output of syllables constructed from observed phoneme combinations. Then the babble begins to reflect the phonemic patterns of the teacher, the quasi-random output is biased towards the perceived syllables in the teacher's utterances. As this continues a word will by chance occur, and this will produce a positive reaction from the teacher, or metaphorically speaking, a 'reward'.[1] Then the word which has been identified is stored in LESA's lexical memory, for future use with possible semantic reference.

2) Progress to date: We have implemented a model of an infant learner / teacher dialog which starts with LESA producing simulated utterances analogous to canonical babbling. This is based on the syllable, taken as one of three types:

- Type 0: vowel alone V, e.g. 'o' as in 'over'
- Type 1: a vowel preceded by one or more consonants, CV, where C can be a single consonant or a cluster, e.g. 'boo', 'blue'
- Type 2: a vowel preceded and succeeded by one or more consonants: CVC e.g 'lift' as in 'lifted'

The syllabic form VC, e.g. 'up', 'at' will be added at a later stage.

Initially LESA produces a sequence of random syllable types, and each type is composed of random phonemes, as found in English[2]. Some syllables are repeated, as in babbling. (However, no use is yet made of the repetition in the learning process)

Parameters such as the length of utterance are based on observed behaviour [19] but semi-arbitrary. For instance, the length of utterance is a random number between 2 and 12. Table I gives an example. For the purpose of exposition letters are used as pseudo-phonemes, and the output is shortened.

utterance 1	
syllable of type 1	jo
syllable of type 2	koth
syllable of type 1	ba
syllable of type 2	nol
utterance 2	
syllable of type 1	po po po po
syllable of type 0	u
syllable of type 0	e
syllable of type 1	le
syllable of type 1	mo mo mo mo mo mo

TABLE I

Sample of untrained output.

Then the teacher will start to talk. Suppose her speech includes "Hello Lesa. Lesa how are you?". In pseudo-phonemes this is represented as an unsegmented stream "h e l o l e s a l e s a h o w a r u". LESA's response will be a semi-random sequence, where the probability of producing each type of syllable 'heard' is increased. There is a bias towards the sounds uttered by the teacher, based on the frequency of occurrence of those sounds, taken in context. See Table II.

utterance 1	
syllable of type 1	he
syllable of type 2	waf waf
syllable of type 2	lew lew lew lew
syllable of type 1	he he he
utterance 2	
syllable of type 0	e
syllable of type 2	lech
syllable of type 1	ho ho ho ho ho ho
syllable of type 0	o
syllable of type 2	wag wag wag wag
syllable of type 1	ho

TABLE II

Sample of output. *After teacher says "h e l o l e s a l e s a h o w a r u" illustrating the increased probability of syllables spoken by the teacher appearing in LESA's output, see 'he' and 'ho' above.*

Learning can be forced by having the teacher repeat the desired items many times. See Table III for an example after the teacher has said "Push the blue one. Touch the red. Pull the box. Red, red, red ..." as represented in pseudo-phonemes.

From the semi-randomized output of LESA the teacher will 'hear' words she wants LESA to learn. Though these are typically subject to segmentation of the phoneme string as explained below, some one-syllable words will occur before segmentation. Suppose the teacher has frequently used the word 'red', then this syllable will occur sooner or later in LESA's output. The teacher recognizes it and makes a response ('reward'). LESA then stores this item in another memory site, and it may be produced in the string of syllables. See Table IV. Note the bold entries in Tables III

[1]The term '*reward*' in the above Figures and our approach refers to feedback abstracted from information on either *communicative success*, or to primary/secondary reinforcing stimuli (real or simulated), including possibly those satisfying a 'social drive' (such as exposure to faces, turn-taking, intersubjective interactions, mutual gaze, skill mastery in a social context, etc.).

[2]A standard Unix random number generator is used. Each item, for instance each consonant, is represented by a number. If there are n items the random number is divided by n and the remainder codes for the item. To introduce a bias the number n is augmented by duplicate items, so the probability of an occurrence is increased.

utterance 1	
syllable of type 0	e
syllable of type 2	boch
syllable of type 1	the the the the the the
syllable of type 0	e
syllable of type 2	pug pug pug pug
syllable of type 1	re
utterance 2	
syllable of type 0	e
syllable of type 1	bo
syllable of type 2	blup
syllable of type 1	pu pu pu pu
syllable of type 2	**red red**

TABLE III

Sample of output. *Output sample after teacher repeatedly says "p u sh th e b l u w o n t u ch th e r e d p u l th e b o x r e d r e d r e d"*

and IV; the interaction between teacher and learner promotes the syllable 'red' to a learnt word.

utterance 1	
syllable of type 1	tu tu tu
syllable of type 0	e
syllable of type 1	the the
syllable of type 1	blu blu blu blu
syllable of type 1	pu pu
word	**red**
syllable of type 2	thesh

TABLE IV

Sample of output. *Output after the word* **red** *has been recognized and 'rewarded' by the teacher*

C. Segmentation Mechanisms

At this stage learning to segment an utterance and extract elements begins. Each time LESA 'hears' a phoneme string it makes an attempt to segment it, using mechanisms described below. These segments constitute new components of the store of perceived phonemic and 'pre-lexical' knowledge, and are available for LESA to use in future productions. With a simulated teacher, LESA's output is checked against a phonetic dictionary (e.g. a subset of the MRC Psycholinguistic Dictionary [20]) to see if it constitutes a proper word or contains a proper word along with other phonemes, or contains more than one word. If it does the teacher will respond positively, when it first occurs, and the segment, a string of phonemes, will be stored. This learnt item (or phoneme transition probabilities occurring within the string) can then be appropriately weighted and produced as part of a future utterance.

With a real human teacher, whose speech would be converted into strings of phonemes, LESA's output would be synthesized speech, and the human would make a subjective judgement to determine reward.

Taking serial learning of short segments as target basic building blocks is consistent with neuroscientific evidence [5] and empirical linguistic observation [21].

In making an attempt to segment a stream of sounds we will take into account the following characteristics of the process in humans.[3] At this stage IDS (infant directed speech) is typically made up of words with few syllables [19].

Salient factors are start and end of utterance markers, of which the latter is the most important. Mothers typically place words on which they want the infant to focus at the end of an utterance. If this word is a noun in English this is usually a grammatical form, but, for instance, in Turkish it leads to ungrammatical productions. In spite of this the practice is widely observed, suggesting that at this stage word segmentation is more important than correct syntax [23].

Another factor is the frequency of certain phonemic combinations. Rare or unobserved sequences tend to indicate a word boundary. Infants of about 11 months have been shown to recognize the phonotactic constraints indicating word boundaries [24].

We will also take account of observed phenomena such as the prevalence in English of the unit known as a 'minimal word' containing a binary foot, composed either of two syllables (e.g. CVCV, where C:consonant and V:vowel) or two moras (e.g. CVC or CVV).

We will investigate methods of integrating these factors to find the placement of word boundaries, via the use of simple, single layer neural networks and information-theoretic methods. Discrete sequential data processed as bigrams or trigrams is typically linearly separable [25].

III. Acquisition of Meaning Through Mediated Physical Interaction

Part of our research focuses on whether it is possible to associate speech and gestural actions of a human with action, visual, proprioceptive, and auditory perceptions of a robot in order to derive 'meaning' for the robot associated with perceived speech patterns.[4] In carrying out this research we take a 'usage-based' view on language acquisition following Tomasello [2] and Bloom [28] and take inspiration from the constructivist work of researchers such as Steels on the emergence of various linguistic capacities in agent communities [29], and also Roy and Pentland [30], Roy [31], and Yu and Ballard [32] in attempting to link perceived speech with object and action perception.

One of the problems faced in an associative approach is that it requires the exploitation of statistical regularities between speech and perception. This implies that many learning episodes would be required. However, although infants are exposed to a high number of ambient speech events and direct feedback learning via their carers (being exposed to as many as 7,000 utterances per day), they actually learn new words with very few presentations. According to

[3] In this current work we do not investigate the role of prosody and intonation, which play a significant role in speech segmentation and enable infants to recognize some organized forms of speech from their earliest months [22].

[4] Following Wittgenstein [26], any derivation of 'meaning' must ultimately be evidenced by appropriate embodied action in language games (cf. [27]).

Bloom [28] 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 (e.g. 50% of the time: see 'milk', hear 'milk' - 50% of the time hear 'milk', see 'cat'- but 'cat' is never associated with 'milk'). Similarly carers rarely provide feedback on actions (Bloom gives the example of the carer arriving home and saying 'Hello baby, whatya been doing today' rather than 'Hello baby, I'm closing the door'). 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 9-month old babies will typically follow the gaze of their mother, follow her pointing gestures and monitor her emotional states. By 1 year old, the infant points on their own and observes the adult's gaze whilst checking if they have changed the adult's attention. If they fail to capture the adult's attention they will alternate between gazing at the object and the adult until they succeed in getting the attention of the adult onto the object. The utterances of the adult together with reinforcement via forms of affective feedback (possibly prosodic features in the adult response) 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 also associated with simple verbal speech patterns of the adult ('see the doggy','hold the bottle'). In our experiments learning will thus rely on the action modalities of the robot together with the robot's ability to share context and referential 'intent' with a human teacher.

A. Software Platform and Architecture

Physical instantiations of a LESA can take humanoid form, such as the iCub [33] or Kaspar2 [34] robots. In carrying out these studies we have modified our existing social learning architecture (ROSSUM [35], [36]), originally implemented on wheeled mobile robots, to the Kaspar2 humanoid robotic platform (see figure 3). The system has been used to learn scaffolded behaviours via directed learning from a human teacher [35] (but not previously linguistic behaviour). We have added additional functionality for face, motion, and colour detection together with a facility for recording the phoneme sequences made by the human teacher, and consider detection of gaze, synchrony, and turn-taking in human-robot interaction. These new modalities are in addition to existing object detection and proprioceptive feedback modalities already present in the architecture (see III-B below).

ROSSUM [36] allows experiments to be carried out where, via the processes of self-imitation and observational imitation, the humanoid robot learns through experiences grounded[5] on its own visual, auditory and sensorimotor feedback the relevant interaction modalities presented by a human tutor. Initially learning would be based on interactions between the human tutor and the robot whereby the tutor reveals to the robot, via speech, deixis, gesture and reference, the various affordances that are available and the relevant effectivities that can be used to exploit these affordances. Here *speaking* is regarded as a particular type of *gesturing*, i.e. motor activity for manipulating the physical or social environment (see [38], [39]). The combination of self- and observational imitation would allow the robot to take both a first and third person perspective in making these discoveries and associate the gestural components with these discoveries.

[5]The idea of 'grounding' refers to relating the meaning of symbols to embodied sensorimotor experience [37].

Fig. 3. ***Learning via Social Interaction.*** *The humanoid robot Kaspar2 is taught how to hold a patterned box.*

B. Progress to Date

The first physical experiments currently being carried out use the University of Hertfordshire's Kaspar2 robot. This is a minimally expressive small humanoid robot with 8 degrees of freedom in the head, and 5 degrees of freedom in each arm. Proprioceptive feedback from each arm is available both when powered (and maintaining a position) and when unpowered. The person interacting with Kaspar2 when unpowered can physically manipulate its arms (for example to make the robot reach for an object) and the proprioceptive actuator readings can be continuously recorded.

Kaspar2 employs two video cameras for its eyes. These are used to obtain images which are processed [40] to yield an additional set of modalities. These include object/pattern recognition, object colour, face detection and motion detection. A number of objects are available for Kaspar2 to interact with. These objects are detected using the ARToolKit system [41]. The objects are pre-learned using ARToolKit, thus Kaspar2 can detect these objects and recognize that they are individual entities in the world, however no other meaning is attached to them. We justify the use of this simplifying step in these early 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 [28]).

The robot also records the directed speech made by the human. The speech pattern of the human is then processed

Fig. 4. ***Sharing Reference with a Teacher in Context.*** *Kaspar2 interacts with the teacher, where both the robot and the human share attention on the coloured objects, a case of rudimentary shared intentional reference.*

to yield a stream of phonemes using a system such as the University of Bielefeld ESMERALDA speech recognition system [42].

C. Experiment II. Grounding Speech - Action and Object Learning

The following experiment is currently underway (experiment II). This is where a a human teacher/'play-pal' interacts with Kaspar2. The human and robot sit on opposite sides of a desk (see figure 4). Kaspar2 initially focuses on the human's face. If the human stays still Kaspar2 will eventually 'get bored' and start to look around. If the human changes his/her face direction (to look at an object) Kaspar2 will broadly follow the line of the human's gaze. During this interaction the human will be encouraged to 'talk' to Kaspar2 and move Kaspar2's arms and hands to push, pull and touch the objects whilst simultaneously explaining the action to Kaspar2. All of Kaspar2's modalities are actively recorded during this period together with the phoneme sequences generated from the analysis of the human speech.

Subsequently the recorded data is analysed firstly to look for clusters of association occurring between the phonemic strings and the 'experiences' (i.e. sensorimotor readings over a time window [43]) in the robot's own modalities. We are currently analysing the data using variants of Crutchfield-Renyi information distance measure [44] as this has been shown to effectively associate sensory data from differing modalities in our previous studies [45], [46], [43]). Subsequently, we will additionally use the database of phoneme strings generated by the experiment in section II above (experiment I). We expect that by using this data the clustering of associations will be biased to the holophrases that the robot is already 'aware of' and to which it will subsequently associate particular classes of objects, actions, or interactions.

D. Assumptions

In carrying out the physical experiments above we make a number of key assumptions. Firstly, that the LESA, embodied as Kaspar2 in this case, is motivated by novelty - thus if the LESA sees a new object or a human face it will continue looking at it. Kaspar2's focus will eventually habituate, after which it will move semi-randomly (randomly but within fixed limits) until it finds a new focus of attention. If it sees the same object again Kaspar2 will fixate on it only if a sufficient time has passed. This 'boredom' threshold is currently set at $\sim$50 seconds. Secondly, the robot is motivated to share the same attention space as the human trainer. Thus if the human decides to look away from Kaspar2 the robot will try and focus in the same general area as the human. We achieve this last step by initially allowing the robot to focus on the human face. If the human head moves a motion detection system will provide a global movement vector, essentially yielding a single integer value describing the angle of movement of the human head. Kaspar2 will then move appropriately with its own head to broadly focus on the same area that the human is looking at.

E. Experiment III: Speaking and Interacting

The output from analysis of the data from experiment II will yield a set of clusters which directly associate the phomenic strings, word sequences, or holophrases with a grounded representation of the robot's perceptions. We intend to use these clusters to re-run experiment II. This time however the robot will 'talk' by pronouncing the associated phoneme strings of any cluster recognized as similar to the current situations (e.g. familiar objects/actions). This phonetic output will be via speech synthesis (using, e.g., the eSpeak system [47]). The aim here is to analyse the effect of human feedback on the robot's learning and on scaffolding futher linguistic competencies. Thus if the human hears the robot say, for example, 'push circle' when pushing a circle object, we would expect the human to provide some reinforcement signal back to the robot, for example 'yes, Kaspar pushes the circle!' or 'No, that's a square!'. This would then allow further analysis to yield valuable information on the nature of the reinforcement.

F. Further Experiments on Negation

The results of continued iterations of experiment III above will further allow us to study the emergence of various forms of negation [48] through the mechanisms of communicative social interaction; indeed, negation has been hypothesized to have been an extremely important qualifier in the emergence of symbolic representation capabilities. Very early in the language development of children negative speech acts emerge, such as the rejective and holophrastic No!, e.g. to refuse certain food or a particular activity. Other functions of negation in early child language include nonexistence, prohibition, denial, inability, failure, ignorance, expressing the violation of a negative norm, and negative inference [49]. The mentioned examples show that the various functions of early negation are not necessarily related to each other and that the term encompasses a set of functions that is remarkably larger in scope than the well known negation of propositions in particular. Which function a particular case of

negation has is obviously highly context-dependent in more than one sense. It depends on the linguistic context on one hand but also on the situational context. An artificial agent that is supposed to appropriate negative humanlike speech acts therefore cannot derive the meaning of these utterances through a simple lexical analysis. It has to consider the situation in which the dialogue takes place (joint attentional frame). Current models either choose the representation of objects [50] or actions [36] as basic representational building blocks. Different functions of negation tend to operate on the other hand more on objects (nonexistence) than on actions (rejection, prohibition), which suggests that the support for certain forms of negation may be rather weak in each of these existing models. Thus, for achieving the emergence of the full range of early negation, ways have to be found to bypass these difficulties.

In experiments currently being designed we intend to consider questions such as:

- Which features must be supported by frameworks for grounded language learning and imitative learning to enable the representation and production of speech acts that involve negation?
- To what degree and in which form must motivation in the robotic platform be modelled for this purpose, as the majority of early negative speech acts are acts of volition and not acts of description?
- Can negation emerge as purely syntactical construction or is it necessary to modify the underlying grounding mechanism?

IV. Discussion, Questions, and Conclusion

Insights of Wittgenstein [26] and Millikan [14], and more constructively Steels [29], [51], suggest that to understand signalling and linguistic behaviour, one needs to take into account usage in its pragmatic embodied social context. The learning of communicative signalling and linguistic systems (at the ontogenic, diachronic, and evolutionary levels) are moreover shaped, not only by details of perception and embodiment (e.g. [52]), but also by details of transmission, sources of error and variability, as well as feedback and repair mechanisms (e.g. [29], [53], [7]).

We have outlined mechanisms for experiments whereby Linguistically Enabled Synthetic Agents (LESAs), are expected to exhibit (1) reinforcing holophrasis and learning of word-level parsing, (2) the grounding of words and lexicon usage frames in action and object learning via physical interactions, and (3) the bootstrapping of simple usage-based proto-grammatical structure via human scaffolding and feedback.

The overall approach is to understand *constructively* what mechanisms could be responsible for the ontogeny of linguistic compentencies. That is, for such a constructive theory of language to be successful it is necessary to build an instantiation that exhibits the phenomenon to be explained, and, moreover, different constructive mechanisms could be assessed against each other by comparing what they actually generate. Preferably these constructivist evaluation test-beds must involve learning in embodied social interactions with humans and physical interactions with rest of the LESA's environment.

Open and challenging research questions in this area include: (1) to what extent can the methods be scaled for human-like acquisition of linguistic abilities?, (2) To what extent does the achievement of one of these stages support the next in autonomous robot learning in social interaction with humans?, (3) Could these methods be extended to apply to human language, say at the level of 3-year old child, or only proto-grammatical approximants thereof?, (4) What 'cognitive' capabilites are necessary for recruitment in the development of human-like linguistic competencies?, (5) Is it necessary to build in universal mechanisms for categorization and generalization, propositional logic, predication, compositional syntax, etc? Can these emerge from more elementary processes, such as Hebbian learning, 'chunking', sequential processing and locality principles or more general cognitive capacities such as perspective taking; action hierarchies; expectation, prospection and refusal?, (6) How can different types of linguistic negation be acquired by a LESA?, (7) To what extent are these mechanisms for the development of linguistic abilities universal, i.e. applicable for any given target natural language? (8) What are appropriate semiotic frameworks for pragmatic acquistion of language usage (e.g. fluid construction grammar [54], dynamic syntax [55])?, (9) To what extent are purported explanations consistent not only with individual ontogeny of linguistic capabilities but also with diachronic (transmission) and evolutionary (phylogenetic) considerations?

The road ahead to understanding language emergence is long and complex, but constructive approaches offer new means and criteria for validating explanations as we progress on this path.

References

[1] B. de Boisson-Bardies, *How Language Comes to Children*. MIT, 1999.

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

[3] P.-Y. Oudeyer, *Self-organization in the Evolution of Speech*. Oxford University Press, 2006.

[4] J. F. Werker and R. C. Tees, "Cross-language Speech Perception: Evidence for Perceptual Reorganization during the First Year of Life," *Infant Behavior and Development*, vol. 7, pp. 49–63, 1984.

[5] F. Pulvermuller, *The Neuroscience of Language*. Cambridge University Press, 2002.

[6] M. M. Vihman and R. A. Depaolis, "The Role of Mimesis in Infant Language Development: Evidence for Phylogeny?" in *The Evolutionary Emergence of Language*, C. Knight, M. Studdert-Kennedy, and J. R. Hurford, Eds. Cambridge University Press, 2000.

[7] A. Wray, "Protolanguage as a holistic system for social interaction," *Language and Communication*, vol. 18, pp. 47–67, 1998.

[8] ——, " ` Needs only´ Analysis in Linguistic Ontogeny and Phylogeny," 2007, pp. 53–70, in [56].

[9] G. Rizzolatti and M. Arbib, "Language within our grasp," *Trends in Neuroscience*, vol. 21 (5), pp. 188–194, 1998.

[10] L. Fadiga, L. Craighero, G. Buccino, and G. Rizzolatti, "Speech listening specifically modulates the the excitability of the tongue muscles: a TMS study," *European Journal of Neuroscience*, vol. 15, pp. 399–402, 2002.

[11] A. Damasio and K. Meyer, "Behind the looking glass," *Nature*, vol. 454, 2008.

[12] M. M. Vihman, "Word learning and the origins of phonological systems," in *Advances in Language Acquisition*, S. Foster-Cohen, Ed. Macmillan, 2008, in press.

[13] R. Borowsky, C. Esopenko, and J. Cummine, "Neural Representations of Visual Words and Objects," *Brain Topography*, vol. 20, pp. 89–96, 2007.

[14] R. G. Millikan, *Varieties of Meaning*. MIT Press, 2004.

[15] SCRIBE, "Spoken Corpus Recordings In British English," http://www.phon.ucl.ac.uk/resource/scribe/scribe-manual.htm, 2004, [last visited 25 July 2008].

[16] P. F. MacNeilage and B. L. Davis, "Evolution of Speech: the Relation between Ontogeny and Phylogeny," in *The Evolutionary Emergence of Language*, C. Knight, M. Studdert-Kennedy, and J. R. Hurford, Eds. Cambridge University Press, 2000.

[17] K. Demuth, "The Prosodic Structure of Early Words," in *Signal to Syntax*, J. Morgan and K. Demuth, Eds. Lawrence Erlbaum, 1996.

[18] J. Blake, *Routes to Child Language*. CUP, 2000.

[19] CHILDES, http://childes.psy.cmu.edu/, 2008, [last visited 25 July 2008].

[20] MRC, "Psycholinguistic database machine usable dictionary," [Last visited 31 October 2008], http://ota.ahds.ac.uk/headers/1054.xml.

[21] C. Lyon and C. L. Nehaniv, "Developing agents that can speak with humans: pointers from the evolution of language," School of Computer Science, University of Hertfordshire, Tech. Rep. 479, June 2008. [Online]. Available: http://homepages.feis.herts.ac.uk/∼comrcml

[22] J. Morgan and K. Demuth, "Signal to syntax: an overview," in *Signal to Syntax*, J. Morgan and K. Demuth, Eds. Lawrence Erlbaum, 1996.

[23] 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*, J. Morgan and K. Demuth, Eds. Lawrence Erlbaum, 1996.

[24] J. Myers, P. W. Jusczyk, D. G. Kemler-Nelson, J. Charles-Luce, A. Woodward, and K. Hirsch-Pasek, "Infants' sensitivity to word boundaries in fluent speech," *Journal of Child Language*, vol. 23, pp. 1–30, 1996.

[25] C. Lyon and R. J. Frank, "Using Single Layer Networks for Discrete, Sequential Data: An Example from Natural Language Processing," *Neural Computing and Applications*, vol. 5, no. 4, pp. 196–214, 1997.

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

[27] 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.

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

[29] L. Steels, "The origins of syntax in visually grounded robotic agents," *Artificial Intelligence*, vol. 103, no. 1-2, pp. 133–156, 1998.

[30] D. Roy and A. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive Science*, vol. 26, pp. 113–146, 2002.

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

[32] C. Yu and D. Ballard, "A multimodal learning interface for grounding spoken language in sensorimotor experience," *ACM Transactions Applied Perception*, vol. 1, pp. 57–80, 2004.

[33] iCub, "RobotCub – An Open Framework for Research in Embodied Cognition," http://www.robotcub.org/, 2004.

[34] M. Blow, K. Dautenhahn, A. Appleby, C. L. Nehaniv, and D. Lee, "The art of designing robot faces - dimensions for human-robot interaction," in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction (HRI 2006), Salt Lake City, Utah, USA, March 2-3, 2006*, M. A. Goodrich, A. C. Schultz, and D. J. Bruemmer, Eds. ACM, 2006, pp. 321–322, see also http://kaspar.feis.herts.ac.uk/.

[35] J. Saunders, C. L. Nehaniv, and K. Dautenhahn, "Teaching robots by moulding behavior and scaffolding the environment." in *Proceedings of the 1st ACM SIGCHI/SIGART Conference on Human-Robot Interaction (HRI 2006), Salt Lake City, Utah, USA, March 2-3, 2006*, M. A. Goodrich, A. C. Schultz, and D. J. Bruemmer, Eds. ACM, 2006, pp. 118–125.

[36] 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

[37] S. Harnad, "The symbol grounding problem," *Physica D.*, vol. 42, pp. 335–346, 1990.

[38] C. L. Nehaniv, K. Dautenhahn, J. Kubacki, M. Haegele, C. Parlitz, and R. 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)*, 2005, pp. 371–377.

[39] 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*, vol. 9, no. 3, pp. 519–550, 2008.

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

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

[42] ESMERALDA, "An environment for statistical model estimation and recognition on arbitrary linear data arrays," University of Bielefeld, Applied Computer Science, https://www.techfak.uni-bielefeld.de/ags/ai/projects/ESMERALDA/index.html, 2005, [last visited 31 July 2008].

[43] N. A. Mirza, "Grounded sensorimotor interaction histories for ontogenetic development in robots," Ph.D. dissertation, Adaptive Systems Research Group, School of Computer Science, University of Hertfordshire, 2008.

[44] J. P. Crutchfield, "Information and its Metric," in *Nonlinear Structures in Physical Systems – Pattern Formation, Chaos and Waves*, L. Lam and H. C. Morris, Eds. Springer Verlag, 1990, pp. 119–130.

[45] L. Å. Olsson, "Information self-structuring for developmental robotics: Organization, adaptation, and integration," Ph.D. dissertation, Adaptive Systems Research Group, School of Computer Science, University of Hertfordshire, 2006.

[46] N. A. Mirza, C. L. Nehaniv, K. Dautenhahn, and R. te Boekhorst, "Grounded sensorimotor interaction histories in an information theoretic metric space for robot ontogeny," *Adaptive Behavior*, vol. 15, no. 2, pp. 167–187, 2007.

[47] eSpeak, http://espeak.sourceforge.net/, 2007, [last visited 31 July 2008].

[48] C. L. Nehaniv, C. Lyon, and A. Cangelosi, "Current Work and Open Problems: A Road-Map for Research into the Emergence of Communication and Language," in *Emergence of Communication and Language*, C. Lyon, C. L. Nehaniv, and A. Cangelosi, Eds. Springer, 2007, pp. 1–27.

[49] S. Choi, "The sematic development of negation: a cross linquistic longitudinal study," *Journal of Child Language*, vol. 15, no. 3, pp. 517–531, 1988.

[50] D. Roy, "Semiotic schemas: A framework for grounding language in action and perception," *Artificial Intelligence*, vol. 167, no. 1-2, pp. 170–205, 2005.

[51] L. Steels, "The Recruitment Theory of Language," in *Emergence of Communication and Language*, C. Lyon, C. L. Nehaniv, and A. Cangelosi, Eds. Springer, 2007, pp. 129–150.

[52] A. Cangelosi and D. Parisi, "The emergence of a 'language' in an evolving population of neural networks," *Connection Science*, vol. 10, no. 2, pp. 83–97, 1998.

[53] K. Smith, H. Brighton, and S. Kirby, "Complex systems in language evolution: the cultural emergence of compositional structure," *Advances in Complex Systems*, vol. 6, no. 4, pp. 537–558, 2003.

[54] L. Steels and P. Wellens, "How grammar emerges to dampen combinatorial search in parsing," in *Symbol Grounding and Beyond*, P. Vogt, Y. Sugita, E. Tuci, and C. Nehaniv, Eds., vol. 4211 of Lecture Notes in Computer Science. Springer, 2006, pp. 76–88.

[55] R. M. Kempson, W. Meyer-Viol, and D. M. Gabbay, *Dynamic Syntax: The Flow of Language Understanding*. Blackwell, 2001.

[56] C. Lyon, C. L. Nehaniv, and A. Cangelosi, Eds., *Emergence of Communication and Language*. Springer, 2007.