What is Needed for a Robot to Acquire Grammar? Some Underlying Primitive Mechanisms for the Synthesis of Linguistic Ability

Caroline Lyon, Yo Sato, Joe Saunders, and Chrystopher L. Nehaniv

Abstract—**A robot that can communicate with humans using natural language will have to acquire a grammatical framework. This paper analyses some crucial underlying mechanisms that are needed in the construction of such a framework. The work is inspired by language acquisition in infants, but it also draws on the emergence of language in evolutionary time and in ontogenic (developmental) time. It focuses on issues arising from the use of real language with all its evolutionary baggage, in contrast to an artificial communication system, and describes approaches to addressing these issues. We can deconstruct grammar to derive underlying primitive mechanisms, including serial processing, segmentation, categorization, compositionality, and forward planning. Implementing these mechanisms are necessary preparatory steps to reconstruct a working syntactic/semantic/pragmatic processor which can handle real language. An overview is given of our own initial experiments in which a robot acquires some basic linguistic capacity via interacting with a human.**

Index Terms—**Cognitive development, developmental robotics, grammar, human–robot interaction , intelligent robots, language.**

I. INTRODUCTION

A ROBOT that can communicate with humans will have to respond to ordinary, natural language—a situation which contrasts with linguistic communication between synthetic agents. In the latter case a system developed from first principles might avoid the vagaries inherent in natural language with its complex syntactic and semantic characteristics. In ongoing work, however, we are developing a system in which a robot will learn to interact linguistically with an untutored, but cooperative human participant. Our work is based on a dialogue between a human speaking naturally to a humanoid robot or synthetic agent, with the eventual aim of developing a system that enables it to acquire language in a data driven manner, without hard-wiring, preprogrammed, syntactic, or semantic linguistic capabilities.

This paper addresses issues of processing actual natural language. Our research differs from other work in the field which typically considers artificially limited vocabularies and may restrict analyses to canonical forms, not encompassing the range

found in unrestricted speech. The robot should be capable of learning to respond to spontaneous utterances of a participant, and to acquire the necessary language competence a grammatical framework will have to be developed. In this paper, we describe some of the mechanisms that underlie grammar learning in general, and that are necessary preparatory steps for grammar development in robots, without claiming to cover all aspects of early language development. We lay out a road map indicating some of the issues that need to be addressed.

Without grammar, a keyword system could be developed, but this would be of limited use. To illustrate this, consider the type of communication used by a chimpanzee that has learned sign language. Terrace *et al.* [1] asked "Can an ape create a sentence?" and answered decisively "no". They report how the celebrated chimpanzee Nim Chimpsky could combine two or even three words appropriately, but an example of a longer production is the following:

Give orange me give eat orange me eat orange give me eat orange give me you.

We can guess at the scenario, but ungrammatical strings of words are hard to understand.

Though semantic, syntactic, pragmatic, and other competencies develop together, we can, to some extent, investigate them separately. In this paper we focus on syntax. Our work is inspired by the acquisition of language in infants, but we also investigate some of the reasons it might have emerged in evolutionary time, and how this might be reflected in the development of an analogous process in a robot. It is beneficial to discuss evolutionary issues when considering issues of the development of linguisitic ability since the former can illuminate the latter. We look at certain underlying mechanisms, some of which are evident in prelinguistic humans, that are likely to be used in the construction of a grammar. This analysis of underlying contributory mechanisms we characterize as *deconstruction*. By examining these various mechanisms underlying grammatical usage we can see possible incremental stages in the emergence of syntax, which may provide some of the key ingredients for the development of linguistic abilities in robots.

In theories on the emergence of language a characteristic of the nativist school is the sudden emergence of highly specialized mental structures. For instance, Chomsky (2005) speaks of "the origins of the faculty of language and its role in the sudden emergence of the human intellectual capacity" [2]. In contrast, we propose an incremental, constructivist approach to the acquisition of language competence, drawing on recent research in neuroscience (e.g., [3]–[5]), in psycholinguistics and infant

language acquisition (e.g., [6]–[8]), and in computerized models (e.g., [9] and [10]).

We bring together these disparate factors in order to plan a route by which a robot or synthetic agent could acquire some competence in natural language, and examine issues arising from the use of real language, not always central to other research (see, for instance, [11] and [12]). Our approach is based on observed characteristics of actual language in use, and we start by describing some of these characteristics. For examples we will sometimes draw on our own work. Simple scenarios such as a blocks world can provide useful stepping stones in the ontogeny of robots that learn to communicate with humans, and we are using such a scenario to investigate preliminary stages in producing an interactive system [13] (and see Section VI).

The paper is organized in the following way. In Section II, we introduce examples of characteristics of natural language that have sometimes been overlooked, and that need to be accommodated. Then in Section III, we look at primitive processing competences that are exapted for language processing in humans. In Section IV, we see how such primitive mechanisms can be deployed and developed at successive levels in language processing: phonetic, morphological, lexical, syntactic, and semantic. In Section V, we examine further syntactic issues that need to be addressed. Section VI gives a brief overview of our current work with the humanoid, child-like robot Kaspar,[1] and Section VII concludes the paper.

II. CHARACTERISTICS OF OBSERVED LANGUAGE

When we examine real data, (dialogues between humans and robots) we find that they differ from the sort of language typically used as a basis of much traditional linguistic analysis. A short example, taken from the teacher-robot dialogue corpus collected in our experiments, is in the Appendix . We find that the language used is quite simple, including nonsentential, but still grammatical fragments. As well as indicative sentences a significant number of imperative and interrogative ones occur.

There is a significant difference in communication emerging within a community of robots (such as in Steels' work [9]) compared to communication between humans and robots. In robot-robot interaction, an optimal "logical" communication system might be expected to emerge, whereas in our work the robot will have to process actual human language, with its accretions of evolutionary baggage [15]. As an example, consider the prevalence of homophones in language. For instance in English in a corpus of about 1 million words, 20 of the 50 most frequently occurring words are homophones [16, p. 19] [2]. It has been suggested that linguistic communication is optimized if a single sound maps onto only one meaning [17]. For instance Oudeyer proposes "For efficient communication, it is better that different words are associated with different meanings." [18]. Nowak says "ambiguity is the loss of communicative capacity that arises if individual sounds are linked to more than

one meaning" and that the absence of word ambiguity is a mark of evolutionary fitness [19, p. 613.] This might be expected in the emergence of a communication system between synthetic agents. However, it is not the case with observed human languages: in English, French, Chinese, Japanese, and other languages, homophones are ubiquitous (e.g., [12], [20] and [21]). Furthermore, commonly used terms can have a spectrum of meanings: in English consider the word "go" which can mean movement or alternatively a form of expressing the future as in "it's going to be cold", or possibly ambiguous terms in between such as "I am going to fetch it".

A. Asymmetric Development

Language acquisition, human or artificial, may be asymmetrical between comprehension and production, in that the ability to comprehend certain linguistic strings does not immediately lead to the ability to produce them. Humans can express the same idea in numerous different ways— language is highly redundant. In a limited blocks world consider a human in dialogue with a robot asking it to take a certain action. The utterances "look at the black box", "have a look at the black box", "now look at the black one" etc., can all express the same concept. Note that semantically key terms may play different syntactic roles. For instance, "look" can be a verb or a noun. If the human is speaking naturally, and has not been tutored to restrict him/herself to certain expressions only, then the robot will have to learn to process any of these alternatives. However, when the robot comes to produce its own utterances, it is not necessary (at least in our initial implementations) to have the ability to produce a variety of expressions in the same way. An example in the Appendix shows how in early experiments a fluent human teacher elicits just single word responses from the robot Kaspar.

Asymmetric development has analogies in humans. For instance, infants can understand much natural speech from adults before they can speak in the same way themselves. Infants recognize grammatical categories and word order early on. Adult utterances that include the usual function words can be better understood than expressions that omit such words, before the stage at which infants typically produce function words themselves [7, pp. 201–209]. In a human infant production typically trails behind perception due to immature articulation, but a robot would not necessarily exhibit this.

III. UNDERLYING MECHANISMS—EXAPTATIONS

In deconstructing grammar we can first look at mechanisms that have been exapted for language processing, which originally evolved for other purposes. Work concerning the acquisition, or ontogeny, of language by a learning agent can be illuminated by an analysis of some primitive prelinguistic mechanisms, which can play key roles in incremental, constructivist learning. Looking at evolutionary factors, we can understand the "illogical" characteristics of natural language and consider how a robot could accommodate them.

We adopt the approach inspired by recent neuroscientific research that there is a dual stream model for language processing: a ventral and a dorsal stream, commonly called the "what" and "how" pathways [13], [22]. The hypothesis is that a ventral stream processes objects and items, using auditory and visual

---

[1]Kaspar is a minimal expressive child-sized humanoid robot developed by the University of Hertfordshire Adaptive Systems Research Group specifically for human–robot interaction. See [14] for design details and rationale.

[2]to, in, for, be, I, by, not, but, are, which, you, there, been, one, we, their, would, so, no, will (not counting *were/where*, which may be homophonous in standard varieties of English, but not in all dialects since either the vowels may differ or the *h* is sometimes pronounced).

stimuli, while the dorsal stream has a procedural role, producing predictive "forward" models. It is proposed that language processing operates a dual system, switching between modes: "a dorsal stream is involved in mapping sound to articulation, and a ventral stream mapping sound to meaning" [23].

A. Serial Processing

For the most primitive animal types to be able to move, motor coordination based on serial processing is necessary. It is also a critical component of language processing. Dominey, Lieberman and Pulvermuller all investigate phenomena associated with serial processing [3]–[5]. Its role is seen in the significance of word order in languages like English, and in the order of morphemes in inflected language. The components of speech are not just a set of items, but ordered sequences. One of the hypothesised functions of the dorsal stream is sequence production.

The fact that we produce and process such well-ordered sequences under real time constraints is demonstrated by the ease with which we disambiguate common homophones, since they are taken in the context of short sequences. Consider "I want two sweets", "I want to go", "I want to too". Serial processing enables homophones to be disambiguated.

A variety of logical connections between two sequences can be expressed by means of simple concatenation and can be seen as underpinned by this basis of serial processing. Objects are qualified: "box" can become "red box". Actions are put into context: "caught" becomes "caught fish". Concatenation can be a mechanism for enhancing the transfer of information.

Furthermore, by using concatenation of the appropriate terms holophrases can be negated or turned into questions. Examples can be found in infant speech: "No bye bye", meaning "don't go away", concatenates "no" to "bye bye". "What Mummy do?" turns a holophrase into a question by concatenating "what". In adult speech a term can be added to an utterance to negate it: "press the bell" and "don't press the bell". In French adding the expression "est-ce que" turns an indicative sentence into a question. Holophrases and phrases can also be concatenated using conjunctions, such as *and*, a form that is often observed in the dialogue between teacher and robot in our experiments.

Different languages have differing patterns of sequential order, but all have a framework within which ordering occurs. Sequential patterns may shift even within variants of the same language, so that in historical time the order of linguistic units seems to emerge through custom and consensus in a community of speakers. Consider the utterance "Isn't it lovely?" from the teacher-robot dialogue corpus. The phrase "isn't it …" is very common, but the noncontracted form "Is not it lovely?", has fallen into disuse. In contrast, the noncontracted form can be heard in expressions that are differently ordered such as "Is it not true that?". From the start we find on the one hand that grammar is not a set of immutable, clear cut rules, but on the other hand characteristic patterns have to be recognized.

B. Categorization

Next we consider primitive sound processing abilities that precede language processing. Categorical perception is the mechanism by which discrete phonetic elements are extracted from an analog acoustic stream. The ability to produce and perceive such elements would extend the range of sounds used in primitive communication.

Categorical perception does not just uncover an existing structure in an analog stream of sound, it develops a structure, which varies from one community to another. Thus, /l/ and /r/ are distinguished in English, but not in Japanese; an extra phoneme between /b/ and /p/ is inserted in one Armenian dialect, though imperceptible to speakers of other dialects and to us. Young infants have the ability to perceive all phonetic distinctions, but this is lost as the child acquires more sensitivity to its ambient language [24]. Oudeyer demonstrates how an analogous process can be modeled by a synthetic self-organizing system, an example of "a form-creating mechanism particularly responsible for shaping living organisms" [25, p. 31].

Categorical perception of phonemes is a specific example of a general propensity for categorization. Templates for phonemes, as well as for syllables and words, are established before a child is one year old. Among other manifestations, the concept of class membership is fundamental to the development of semantic and syntactic organization of language. This is reflected in neuronal structures, where it is found that different areas are activated by different types of words: those that are associated with actions and those that are associated with objects [5]. However, as discussed below, this distinction does not entirely match standard syntactic part-of-speech categories.

Work on modelling symbol grounding, inspired by actual neuronal processing, has made progress [26]. This is typically based on the hypothesis that symbols are grounded in internal categorical representations (semantic meaning) and have relationships with other symbols (syntactic structure). This approach has been adopted in our work (Section VI).

IV. DEPLOYMENT AND DEVELOPMENT OF PRIMITIVE MECHANISMS

We now point to the likelihood that the primitive mechanisms of serial processing and categorization are recruited at successive levels in language processing. The rationale is that, based on such mechanisms, language specific capabilities can gradually emerge at phonetic, morphological, lexical, syntactic, and semantic levels.

A. Prediction— Forward Planning

A universal characteristic of language is that it is a phonemic system based on the syllable. It is clear that the perception and production of syllables require the combination of the ability to categorize with a serial processing mechanism. For the production of syllables, there is a need for forward planning that anticipates future output. For example, in English, the phoneme /d/ requires different lip positions before "do" and "day", or /t/ before "to" and "tea". At the start of producing these syllables, the speaker must have adopted the appropriate lip positions (which a finger against your lips as you speak will demonstrate). As soon as a small child can say "Dada" and "doggy", he shows he can plan ahead in speech production. This is an example of serial processing: the components of a short sequence are in the mind before the start of the sequence is uttered.

Such a capacity for forward planning also anticipates the way syntactic structures may be acquired and generated at a later stage. The importance of this capacity has been emphasized in the account of the "incremental" parsing of phrase structure [27], [28]. According to this as soon as, and each time, the hearer encounters a single word successively in an utterance he forms or adjusts his anticipation as to the eventual completed "goal" phrase or sentence. Given this tendency an incremental parser is an appropriate model to use for the natural language interaction between human and robot, such as that described in [29].

Such forward planning also points to the similarity in processing between perception and production. For, if hearing part of a linguistic unit triggers prediction of what might be eventually said, it is very much like its production, in the sense that both procedures serially realize it with an anticipated goal in mind. This relationship may be found in lower level processing such as the perception/production of syllables.

B. Segmentation

Information-theoretic methods can show that segmenting speech into appropriate chunks makes communication more efficient [30]. From the point of view of the perceiver, when a speaker produces a syllable, it will activate some of the same neural structures, as if the hearer is about to produce that syllable [31]. When a string of phonemes is produced, the hearer needs to be able to break up the utterance and impose the appropriate structure. We note the observed phenomenon that prelinguistic infants perceive patterns in meaningless strings of syllables and are capable of segmentation [6, p. 1034].

The ability to segment a stream of sounds can be applied to the detection of linguistic units using a number of different mechanisms, described in Section VI. A linguistic unit is a sequence of one or more syllables that could constitute a part of a word, a word, or more than one word, a holophrase, or grammatical phrase. In an interactive dialogue with a mature speaker, a prelinguistic infant babbles its syllabic productions, though probably without meaning, which become biased towards the sounds heard in ambient language or spoken by the teacher. The teacher might reinforce any chance syllabic output from the infant that resembled a proper linguistic unit. A preliminary stage in learning the meaning of utterances is followed by the concatenation of linguistic units to describe objects or events, ask questions, issue requests, to express feelings, and to manipulate the world by influencing the behavior of others through language. At an early stage, an infant may not process all of a perceived utterance, but only part of it.

It has also been suggested that distributional analysis can contribute to segmentatin in that frequent syllables and syllable strings are identified as words. However, infants typically learn to produce content words prior to function words, despite the relative frequency of the latter over the former. Therefore, some semantic saliency may play a part in learning to recognize words.

C. Compositionality and a Hierarchical Grammar

As in the production and perception of phonemes and syllables, linguistic units such as words can be combined in various ways, conforming to overall acceptable patterns.

Thus, the infant may hear phrases with shared words, such as "small boy", "big boy", and "big shoes", and gradually induce the fact that "small" and "big", or "boy" and "shoes" belong to the same respective categories. The same type of linguistic units are then interchanged and lead to "small shoes", a novel form of compositional structure.

This characteristic is used in unsupervised, alignment based learning parsers, because if parts of sentences can be substituted for each other, then these constituents are of the same type [32]. The alignment based learning approach, which has had some success, differs from ours in that it analyses only complete sentences.

Steels points out that compositional structure makes for reduced lexicon size compared to a language in which each phrase has its own representation, and thus, to reduced computational load. He says "the first purpose of grammar is to reduce the number of variables in a decoded meaning structure and hence reduce the computational complexity of its interpretation" [33, Sec. 3].

However, there is another, possibly primary, reason for the emergence of compositional structure: items with related *meanings* have related representations. We need to model semantic cohesion. While a three-word phrase such as "the red ball" could have a number of possible interpretations, it is not merely the accepted compositional pattern that narrows them down to the probable combination, but also semantic cohesion. In order to recognize the word "red" in this utterance, the hearer must have been previously exposed to the same word in a context most probably different from the present one, perhaps "the red box". The semantic recognition that the red ball and the red box have something in common in meaning should restrict the assignment of interpretations, and hence communication is enhanced by compositional structure.

Simple compositional structures may include the following.

- Modification: as in "red ball" or "small boy", concatenating a unit that qualifies another unit.
- Predication: the attribution of a property to an entity, such as "the moon is white".[3]
- Conjunction: the concatenation of two similar linguistic elements, such as "and" to join two adjectives or two phrases or two sentences.
- Interrogation: converting an assertion to a question by adding an interrogatory marker, such as the French "est-ce que".
- Negation: adding a negating term to counter an utterance, applied to objects ("not food"), actions ("not touch") or descriptors ("not good"). Negation can be verbal, non-verbal (e.g., gestural) or both. It is more informative to negate an existing utterance than to produce a new unrelated one.

Compositional structure is the basis for a hierarchical grammar: two or more words can be combined to act as a single unit, for instance *adj noun → nominal*. Then this compound unit can be combined with another word to constitute a new element, e.g., with an article to constitute a noun phrase, such as "a black cat". As this process is repeated a hierarchical structure is developed from the bottom up.

---

[3]This is actually the basis for all propositional semantics, and its emergence is by no means understood as yet (neither in ontogeny nor evolution)—perhaps it arises via the grammaticalization of topic-comment associative utterances.

V. FURTHER CHARACTERISTICS OF A FULL SYNTAX

At this stage we have not mentioned the word "sentence" which may involve recursion or distant dependencies, properties considered hallmarks of natural language [34]. Algebraically, it is only a small step from the word combination described above to achieve these two properties: a compound unit itself becoming a constituent of another, higher level, compound unit. A real issue of language acquisition is, however, how the learning agent discovers this key additional component in an algebraically less clear-cut set of data, and we sketch below phenomena that need to be captured.

The extraction of keywords like "look" and "black" can be implemented in a relatively straightforward manner. But as a model of language acquisition by an agent, merely collecting words will soon be inadequate even in the blocks world. For instance, as soon as expressions like "put the black ring on the red box" occur spatial relationships must be modelled, indicating which item is above or below the other.

A. Sentences, Grammatical Fragments, and Rules

It has been shown that communication is more efficient if an utterance is divided up into appropriate segments which correspond to syntactic components. The ease with which an utterance can be decoded is related to the entropy of the sequence, and the entropy declines if a string of words is segmented. Such segments are sentences and also subsentential grammatical elements [30].

Looking at the actual language used in our human-robot experiments (and in dialogues with infants as collected in the CHILDES corpus [35]) we note that the canonical sentence form of *subject-predicate* or *noun phrase-verb phrase* is not always the most common. There are numerous imperatives: "Look at the moon shape", and questions: "Can you see the star?". Indicative sentences frequently start with a presubject word or phrase: "Kaspar, here is a circle". Others are concatenated with a conjunction. As in most spontaneous speech there are well formed utterances that are not sentences: "That's a picture of a heart. A picture of a heart." A grammar will have to handle all these variations, as described in [29].

When we look at real language we quickly see that rules can be considered as prototypes that are frequently amended, adapted, or ignored. Much of the sort of language that we can expect to be spoken to our robot will be based on a consensus of usage rather than on rules. In developing communication between humans and robots, we will need a grammar that encompasses the different forms of acceptable usage. Thus, an approach which considers that "a grammar can be seen as a rule system that divides [...] sentences into two subsets, grammatical and ungrammatical" [36] will be too narrow for our purposes.

To illustrate this, consider the previous example of the question "Isn't it?", or the usage of the following words which are likely to occur in spontaneous dialogue: *all, both, either, some, any*

- All of the boys, both of the boys, either of the boys, some of the boys, any of the boys.
- All the boys, both the boys, *either the boys, *some the boys, *any the boys.

- All we have is …, *both we have is …, *either we have is …, *some we have is …, etc.

To try to put these "construction islands" [8] into rules may mean creating a set of rules for each word. Though this approach has been implemented with a degree of success, for instance with a Link grammar [37], it cannot cater for grammatical fragments, and typically produces numerous alternative candidate parses.

B. Anaphoric Reference and Distant Dependencies

Another significant issue is the representation of anaphoric references, which occur frequently in interactions between humans and humanoid robots. For example, the words "it" or "that" need to be linked to the earlier specified noun which they represent, as in:

"Can you see the circle? It's black, isn't it? That's a black circle."

Steels' approach in fluid construction grammar (FCG) is based on semantic and syntactic "poles" or processors in tandem which will facilitate the representation of anaphoric references. His approach could also be useful in that "[The] parser and the constraint network are 'fluid' in the sense that they attempt to arrive at an interpretation even if there are unknown words" [38, Sec. 2].

There is also the issue of distant dependencies with feature percolation within sentences. For a head structured grammar, such as English, conventions on agreement may apply to words that are at a distance from each other.

For instance, in: "The stars on the box are painted white," the verb "are" is plural to match the head of the subject "stars", rather than the adjacent single noun phrase "the box". This property has been a centre of attention in various grammatical frameworks, but from a purely practical point of view the semantic interpretation may be clear without this detailed parsing. What seems to be required in learning theory, but relatively less explored in grammar, are semantic and pragmatic facilitation effects in such distant dependencies.

C. Transferable Reference —The Use of Pronouns

A child's understanding that a pronoun like "I" means Mummy when Mummy uses it, but means the child when the child uses it about herself has to be learned. In early talk, the child will often refer to herself or her mother by name when an adult would use a pronoun: "Mummy's cake" rather than "your cake". One way to learn the use of pronouns is from ambient speech, not directed at the child, so the child will hear the same pronoun, such as "I" or "you" used to refer to different people. In our work we do not (yet) touch on this issue, but restrict our experiments to one-to-one dialogue between the synthetic agent or robot and a teacher.

D. Embedding and Recursion

Recursion, where an element is defined in terms of an element like itself, is a key characteristic of human language, but is not prominent in our corpus of simple human-robot speech. However, it occurs on a small scale in embedded phrases such as in "a large black circle with a small white circle in the middle", decomposing into

nominal → nominal preposition nominal

and so on. We cannot, as yet, provide a satisfactory account in

detail of how a robot may learn recursion: embedding of structures within structures is what is required. This might arise via concatenation to express a particular semantic relation which is later subject to grammaticalization yielding embedded constructions (cf. [8]). We investigate how the robot can come to take such a phrase to be recursive, as a stepping stone towards a full recursive syntax. It needs little imagination to see how useful sentence recursion would be from the earliest times. For example, in some primitive hunting community a member of a group planning an ambush might want to pronounce:

"the chief says that we must wait here"

decomposing into

sentence → sentence "that" sentence.

Note that when Chomsky and his school discuss recursion they are typically considering "the abstract linguistic computational system alone" [34]. Such a computational system can yield infinite recursion, which contrasts with recursion in the real world where there are obvious practical limitations.

E. Revisiting Traditional Practices

In human–robot interaction in a blocks world scenario, and elsewhere, the common distinction between content words and function words may need amending. Prepositions belonging as they do to a closed class are considered function words, but we may want to consider them as content words since they carry significant meaning in our scenarios. Consider Steels' example of decoding "The ball that hit the box next to the green cube" where he classes *nextto* (sic) as a content word.

We also need to examine our categories of nouns and verbs in the light of recent results from neuroscience. Pulvermuller [5] reports on neurophysiological investigations into noun and verb processing, showing that different areas of the brain process object words and action words, but action words include nouns associated with actions. He gives the examples of "whale" commonly understood as a visual object and "fork" usually associated with the action of eating. "There was no difference in the topography of brain responses between action verbs and nouns for whom strong action associations were reported."(ibid, p 61). And as mentioned in Section II-A, a word like "look" can be used as a verb or a noun with the same conceptual meaning.

VI. Current Work

This paper examines mechanisms underlying grammar which need to be taken into account in developing a robot that can acquire language competence. Experiments with our own implementation of such a robot have started [13], and while this paper is an attempt to look at the lie of the land ahead, it is useful to describe our ongoing experimental research and immediate future plans. Our work is based on a dialogue between a human speaking naturally to a humanoid robot or synthetic agent, with the eventual aim of developing a system that enables it to acquire language, taking a phased approach. We start with the assumptions that our robot can learn, hear speech (or simulated speech to an agent), and take turns in dialogue. Auditory input is represented as a stream of phonemes.

One set of experiments, analogous to dorsal processing, models the early stages of language acquisition, moving from infant babbling to perception and production of words and holophrases. Syllable based babbling is an example of basic serial processing of short sequences of phonemes. Initial canonical babbling, random production of syllables, becomes biased towards the syllables of the carer, and eventually a word, or holophrase with a word embedded, is produced. At this stage the synthetic agent has not developed the capability of associating words or holophrases with meaning, but can respond to a metaphorical "reward" from a teacher by reinforcing appropriate behaviour. This stage reflects the observation, discussed in Section IV, that the prelinguistic infant already uses some of the mechanisms needed for a mature syntax when he attends to the talk of a carer or to ambient speech of others around, and produces babbling, syllable based sounds [39, chapter 5].

At the next level the young infant or synthetic agent can segment perceived streams of syllables even when they do not have any meaning, and "[t]he first (presymbolic, prereferential, context-limited) words produced reflect a match between the child's babbling patterns and adult patterns produced in a meaningful context" [40, p. 136].

Apart from interaction with a human teacher, the infant or robot may use other mechanisms to determine word and phrase boundaries. These may include phonotactic patterns, prosodic information and the placement of words at the start or end of a break in an utterance [41], [42]. Phonotactic constraints mean that less probable transitions from one syllable to another indicate possible word boundaries [43]. Prosodic information, from tone, pace, and stress patterns, plays a significant role in determining boundaries of linguistic elements (but in practice it can be hard to capture tone and stress automatically). It has been noted that in speaking to infants carers often put salient words at the end of an utterance. In English this is usually grammatical, but the practice is also observed in languages where the result may not be grammatical, such as Turkish. In experiments carried out by our group on human-robot interaction this tendency is found to be useful as the robot starts to learn the meaning of words [13].

In further experiments, analogous to ventral processing, the robot, a humanoid small child Kaspar (see figure) learns to attach meanings to words and holophrases, through a process of joint reference with its human teacher. It can see a limited local scene and respond to certain weighted combinations of sensory perceptions, drawing on heuristics derived from the experience of human infants in language acquisition [44]. This is an example of the underlying propensity to categorize. At our present early state of development, Kaspar will respond with single words, which depend on its history of interaction with the particular partner, and are not necessarily initially the right ones according to the teacher's usage. However, in the course of several sessions of interaction, associations between the presence of the objects in sensorimotor streams and uttered words do correspond correctly according to the teacher's naming of the objects, thus grounding the robot's linguistic reference to the objects, which it subsequently expresses in its own utterances. An example is given in the Appendix.

When the infant produces his first simple words, he is already displaying the ability to plan ahead. As the infant moves on to the acquisition of words with meanings, so does our robot. Experiments have shown that humans use language similar in

Fig. 1. The humanoid robot Kaspar II, analogous to a young child, working with a teacher.

many respects to converse with a child-like humanoid robot as they might use with a real child [44]. A rudimentary form of shared attention is exploited, and the robot can associate the teacher's words and its perception of different shapes together with other visual and auditory perceptions, and actuator proprioceptions [44].

At the third stage, we plan to start testing a module for initial grammar learning, which is presently being implemented. The overall approach hinges on the observation that when the infant produces his first simple words associated with rudimentary meaning or reference, he is already displaying the ability to plan ahead characteristic of the syntax of the mature speaker. This suggests a "bootstrapping" approach to syntax learning, based on the meanings of recognized words, as described in [45].

Last, we need to model the subsequent learning process applied to the preliminary, incomplete grammar. In processing the auditory input (the teacher's speech) simple mechanisms, such as taking syllable(s) at the end of an utterance as a salient word, initially works well; but the robot outgrows this approach, so another process is required. At this point, a predictive grammar, a forward model is needed to filter out multiple candidates and capture meaningful utterances. The grammar outlined in [29] may provide a starting point.

VII. CONCLUSION

If a linguistically enabled robot can construct a grammar to facilitate communication with humans, it will need to be based on primitive mechanisms, such as serial processing and categorization, which underlie grammatical systems. These mechanisms seem to be exaptations of processes originally developed for other purposes, but then recruited for linguistic communication. Other factors also come into play, for instance, the segmentation of an utterance into appropriate chunks makes communication more efficient. Basic mechanisms then combine with other factors and are deployed at a higher level. For instance, the ability to categorize combined with serial processing enables compositional structuring, and leads to forward planning

of linguistic components. Compositionality leads to a hierarchical structure. However, though this is a short step in a logical system, it is not yet clear how a robot can learn this.

We see that the actual language used in interactions between humans and humanoid robots does not always match the canonical forms on which some analyses are based. The first requirement is to observe the actual language used. From this we can deconstruct the grammar in order to derive underlying primitive mechanisms, and then aim to reconstruct a working syntactic-semantic-pragramtic processing mechanism that can handle real language.

In deconstructing a grammar, we touch on the acquisition of language in evolutionary time, in historical time, and in the lifetime of an individual. We suggest that in developing a linguistically enabled robot we need to draw on observed features on each of these scales.

APPENDIX

Two examples follow of speech taken from the teacher–robot dialogue corpus collected during our experiments, in which a participant shows the child-like robot Kaspar some objects and asks him questions about them.

Teacher's Speech Before Kaspar has Learned to Respond:

hello kaspar
what have you got here
ah got a box
some shapes on it what shapes this got
circle shape with black circle and a white dot
can you see that
and if we move the box round weve got other shapes
and what shape have we got here
this is another black and white shape
with a moon on it
can you see thats got a moon and
if we turn the box round again
got another shape this is a star shape its
a black type of sun actually
its got a lot of little triangles on it
and a circle and a big circle in the middle
with loads of triangles and it makes a sun shape
and whats on this side of the box
and weve got a heart can you see that

Dialogue When Kaspar is Beginning to Learn the Names of Shapes:

Kaspar's words are marked with angle brackets.
hi kaspar look whats that
>> square
yes and that
>> moon
yes and that
its a sun
>> sun
yes
>> kaspar

and this
>> heart
and this sign
>> triangle
yes brilliant and this one
>> shape
a circle
>> circle
yes
>> square
square thats a square sign
whats that one
its a heart
>> heart

End of sample dialogue.

REFERENCES

[1] H. S. Terrace, L. A. Pettito, R. J. Sanders, and T. G. Bever, "Can an ape create a sentence?," *Science*, vol. 206, pp. 891–902, 1979.

[2] N. Chomsky, "Three factors in language design," *Ling. Inq.*, vol. 36, no. 1, pp. 1–22, 2005.

[3] P. F. Dominey, M. Hoen, J.-M. Blanc, and T. Lelekov-Boissard, "Neurological basis of language: Evidence from simulation, aphasia and ERP studies," *Brain Lang.*, vol. 86, pp. 207–225, 2003.

[4] P. Lieberman, *Human Language and our Reptilian Brain*. Cambridge, MA: Harvard Univ. Press, 2000.

[5] F. Pulvermuller, *The Neuroscience of Language*. Cambridge, U.K.: Cambridge Univ. Press, 2002.

[6] A. Fernald and V. A. Marchman, "Language learning in infancy," in *Handbook of Psycholinguistics*, M. J. Traxler and M. A. Gernsbacher, Eds., 2nd ed. New York: Academic, 2006, pp. 1027–1071.

[7] B. de Boisson Bardies, *How Language Comes to Children*. Cambridge, MA: MIT, 1999.

[8] M. Tomasello, *Constructing a Language*. Camrbidge, MA: Harvard Univ. Press, 2003.

[9] L. Steels, *The Talking Heads Experiment*. New York: VUB, 1999.

[10] A. Cangelosi and D. Parisi, *Simulating the Evolution of language*. New York: Springer-Verlag, 2001.

[11] G. Sampson, *Educating Eve: The language Instinct Debate*. London, U.K.: Cassell, 1997.

[12] C. Lyon, C. L. Nehaniv, S. Warren, and J. Baillie, "Homophony and disambiguation through sequential processes in the evolution of language," in *New Frontiers in Artificial Intelligence*. New York: Springer-Verlag, 2007, vol. 3609, Springer Lecture Notes in Computer Science, pp. 315–324.

[13] J. Saunders, C. Lyon, C. L. Nehaniv, K. Dautenhahn, and F. Forster, "A constructivist approach to robot language learning via simulated babbling and holophrase extraction," in *Proc. IEEE ALife Conf.*, Chicago, IL, 2009, pp. 13–20.

[14] K. Dautenhahn, C. L. Nehaniv, M. L. Walters, B. Robins, H. Kose-Bagci, N. A. Mirza, and M. Blow, "Kaspar – A minimally expressive humanoid robot for human-robot interaction research," *Applied Bionics and Biomechanics, Special Issue on Humanoid Robots*, to be published.

[15] C. Lyon and J. Saunders, "Investigating the basis for conversation between a human and robot," in *Proc. Int. Conf. Epig. Robot.*, Venice, Italy, to be published.

[16] S. Johansson and K. Hofland, *Frequency Analysis of English Vocabulary and Grammar*. London, U.K.: Clarendon, 1989.

[17] M. Redford, C. C. Chen, and R. Mikkulainen, "Constrained emergence of universals and variation in syllable systems," *Lang. Speech*, vol. 44, no. 1, pp. 27–56, 2001.

[18] P.-Y. Oudeyer, "Language evolution as a Darwinian process: computational studies," *Cogn. Proc.*, vol. 8, no. 1, pp. 21–35, 2007.

[19] M. A. Nowak, N. L. Komaraova, and P. Niyogi, "Computational and evolutionary aspects of language," *Nature*, vol. 417, pp. 611–617, 2002.

[20] J. Ke, F. Wang, and C. Coupe, "The rise and fall of homophones: A window to language evolution," in *Proc. 4th Int. Conf. Evol. Lang.*, Harvard, MA, 2002.

[21] L. Lamel, M. Adda-Decker, and J.-L. Gauvain, "Issues in large vocabulary, multi-lingual speech recognition," in *Proc. Eurospeech*, 1995, pp. 185–188.

[22] G. Hickok and D. Poeppel, "Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language," *Cognition*, vol. 92, pp. 67–99, 2004.

[23] D. Saur, B. W. Kreher, S. Schnell, D. Kummerer, P. Kellmeyer, M.-S. Vry, R. Umarova, M. Musso, V. Glaucher, S. Abel, W. Huber, M. Rijntjes, J. Hennig, and C. Weiller, "Ventral and dorsal pathways for language," in *Proc. Nat. Acad. Sci.*, 2008, vol. 105 (46), pp. 18 035–18 040.

[24] J. F. Werker and R. C. Tees, "Cross-language speech perception: Evidence for perceptual reorganization during the first year of life," *Inf. Behav. Develop.*, vol. 7, pp. 49–63, 1984.

[25] P.-Y. Oudeyer, *Self-Organization in the Evolution of Speech*. Oxford, U.K.: Oxford Univ. Press, 2006.

[26] A. Cangelosi, "Approaches to grounding symbols in perceptual and sensorimotor categories," in *Categorization in Cognitive Science*, H. Cohen and C. Lefebvre, Eds. Amsterdam, The Netherlands: Elsevier, 2005, pp. 719–737.

[27] F. Costa, P. Frasconi, V. Lombardo, and G. Soda, "Towards incremental parsing of natural language using recusive neural networks," *Appl. Intell.*, vol. 19, no. 1-2, pp. 9–25, 2003.

[28] P. Sturt and M. Crocker, "Monotonic syntactic processing: A cross-linguistic study of attachment and reanalysis," *Lang. Cogn. Proc.*, vol. 11, pp. 448–494, 1996.

[29] R. Kempson, E. Gregoromichelaki, and Y. Sato, "Incrementality speaker/hearer switching and the disambiguation challenge," in *Proc. ACL Workshop Seman. Rep. Spoken Lang.*, Athens, Greece, 2009, pp. 74–81.

[30] C. Lyon, B. Dickerson, and C. L. Nehaniv, "The segmentation of speech and its implications for the emergence of language structure," *Evol. Comm.*, vol. 4, no. 2, pp. 161–182, 2003.

[31] M. Arbib, "From monkey-like action recognition to human language: An evolutionary framework for neurolinguistics," *Behav. Brain Sci.*, vol. 28, no. 02, pp. 105–124, 2005.

[32] M. van Zaanen, "Bootstrapping syntax and recursion using alignment-based learning," in *Proc. 17th Int. Conf. Mach. Learn.*, 2001, pp. 1063–1070.

[33] L. Steels, "What triggers the emergence of grammar?," in *Proc. 2nd Int. Symp. Emerg. Evol. Ling. Comm.*, Hertfordshire, U.K., 2005, pp. 143–150.

[34] M. D. Hauser, N. Chomsky, and W. T. Fitch, "The faculty of language," *Science*, vol. 198, pp. 1569–1579, 2002.

[35] *CHILDES: Child Language Data Exchange System*, , 1984–2009 [Online]. Available: http://childes.psy.cmu.edu

[36] N. Komarova and M. Nowak, "Population dynamics of grammar acquisition," in *Simulating the Evolution of Language*, A. Cangelosi and D. Parisi, Eds. New York: Springer-Verlag, 2002, pp. 149–163.

[37] D. Sleator, D. Temperly, and J. Lafferty, Link Grammar Carnegie Mellon University, 2005 [Online]. Available: http://www.link.cs.cmu.edu/link/

[38] L. Steels and P. Wellens, "How grammar emerges to dampen combinatorial search in parsing," in *Proc. 3rd Int. Symp. Emerg. Evol. Ling. Comm.*, Rome, Italy, 2006, pp. 76–88.

[39] P. Juszcyk, "How infants begin to extract words from speech," *Trends Cogn. Sci.*, vol. 3, pp. 323–328, 1999.

[40] 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, MA: Cambridge Univ. Press, 2000, pp. 130–145.

[41] 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. New York: Erlbaum, 1996, pp. 117–134.

[42] 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," *J. Child Lang.*, vol. 23, pp. 1–30, 1996.

[43] J. Saffran, R. Aslin, and E. Newport, "Statistical learning by 8-months-olds," *Science*, vol. 274, no. 5294, pp. 1926–1928, 1996.

[44] J. Saunders, C. L. Nehaniv, and C. Lyon, *Robot Learning of Object Semantics from the Unrestricted Speech of a Human Tutor*, 2009, submitted for publication.

[45] Y. Sato and J. Saunders, "Semantic bootstrapping in embodied robots," in *Evolang 8*, 2010, submitted.

Caroline Lyon received the B.A. degree in philosophy, politics, and economics from Oxford University, Oxford, U.K., in 1963. She received the M.Sc. and Ph.D. degrees in computer science from the University of Hertfordshire, Hatfield, U.K., in 1988 and 1993, respectively.

She is currently a Visiting Research Fellow at the University of Hertfordshire, Hatfield, U.K. Her research interests focus on early cognitive developments and their analogs in artificial inteligence, but also include computational linguistics, natural language processing, speech recognition, and the evolution of language.

Dr. Lyon coedited the book, *Emergence of Communication and Language* (Springer) in 2007.

Yo Sato received the M.Sc. and Ph.D. degrees in computational linguistics from King's College London, London, U.K., in 2002 and 2007, respectively.

He was a Postdoctoral Research Fellow at Queen Mary, University of London, U.K. He is currently a computational linguist and is working as a Research Fellow at the University of Hertfordshire, Hatfield, U.K., on the language learning algorithms for cognitive robots in a European project. His main interests include grammar formalisms, computational syntax and semantics, parsing algorithms, dialogue processing, and unsupervised grammar induction.

Joe Saunders received the M.Sc. and Ph.D. degrees in computer science from the University of Hertfordshire, Hatfield, U.K., in 2003 and 2007, respectively.

He has published research articles on robot learning, imitation, and human–robot interaction, and is currently involved in several European robotics projects. Prior to 2003, he directed large software development projects for a number of international investment banks. His research interests include social robotics, language, and imitation in artificial systems.

Chrystopher L. Nehaniv received the B.Sc. degree from the University of Michigan, Ann Arbor, in 1987, and the Ph.D. degree in mathematics from the University of California, Berkeley, in 1992.

He is currently a Research Professor of Mathematical Evolutionary Computer Sciences at the University of Hertfordshire, Hatfield, U.K., where he founded the Adaptive Systems Research Group together with Kerstin Dautenhahn in 2000. His main interests include interactive systems, complex adaptive systems, artificial life, evolvability, emergence of language, and communication in embodied robots interacting with humans, as well as, abstract algebra (theory of semigroups, groups, and sequential machines) and its applications, especially, algebraic biology.

Dr. Nehaniv is also an Associate Editor of the journals *BioSystems* and *Interaction Studies*.