An Integrated Three-Stage Model Towards Grammar Acquisition

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

***Abstract*—This paper presents a three-stage model of language acquisition that integrates phonological, semantic and syntactic aspects of language learning. With the assumption that these three stages arise roughly in sequence, we test the model using the experimental methodology of cognitive robotics, where an emphasis is placed on situating the robot in a realistic, interactive environment. The first, phonological stage consists in learning sound patterns that are likely to correspond to words. The second stage concerns word-denotation association, which relies not only on sensory input but also on the learner's speech output in 'dialogue'. The data thus gathered allows us to invoke *semantic bootstrapping* in the third, grammar induction stage, where sets of words are mapped with simple logical types. We have started implementing the model and report here on the initial results of the human-robot interaction experiments we conducted.**

I. Introduction

This paper presents a computational model, implemented as a prototype in a simple human-robot interaction (HRI) scenario, of language acquisition analogous to infants' development processes. Learning takes place in three stages on different aspects roughly in sequence, though overlapping: word sound form detection, word-denotation association and grammar induction. To test the model, we use the methodology of cognitive robotics in an HRI experiment. After presenting the model, we describe the HRI experiments we have conducted and report on some results.

Our underlying rationale is to place the burden of learning on phonological processing and, above all, semantic understanding, in a move towards a performance-based theory of language (e.g. [1], [2]). This contrasts with the need for innate syntax, such as has long been advocated in theoretical linguistics. Our model shows affinity to the concept of *semantic bootstrapping* dating back to [3], though we redress the balance from syntax towards semantics. The basic idea is that the learning agent starts with sound pattern recognition, proceeds then to two stages in the realm of meaning, first word meaning and then sentence meaning, acquiring rudimentary syntax in so doing.

For grammar induction, a central focus in the present work, we primarily invoke a mechanism of formal semantics [4], albeit in an underspecified manner to reflect the rudimentary nature of infants' semantic understanding (discussed in Section III-C). Our original contribution will thus be twofold: the use of the cognitive robotics methodology on one hand to integrate these three stages, and the theoretical machinery of formal semantics on the other to induce grammar, in an attempt to illuminate the empirically pivotal issue of language acquisition. We believe that combining the techniques developed in these two domains, which has not been traditionally explored, will shed new light on the domain of language acquisition research.

The paper is organised as follows. After an overview of the current state of related research in the next section, Section III describes our model with its three separate subsections devoted to the three stages. Section IV describes our experiment and reports on some results and the discussion thereof, and we conclude with some remarks on future work.

II. Background

While it is extremely difficult in the empirical developmental studies on language acquisition to disentangle the different aspects of language —phonological, semantic, pragmatic and syntactic— it is useful, as well as feasible, to implement and test a computational model that differentiates distinct stages of language acquisition. Our model proposes a three stage process: word form detection, word-denotation association and grammar induction. Empirical evidence suggests that infants recognise and can produce frequently coherent sound sequences without meaning, prior to speaking [5]. Their earliest speech productions then are typically single-word utterances where the words are generally restricted to content words, later progressing to multi-word utterances. Furthermore, infants' initial multi-word utterances frequently come in the form of two-word combinations, including what Braine [6] called 'pivot schemata', such as '*Mummy gone*', '*More milk*' or '*All clean*', which distinctly look like a predicate-argument structure, albeit incomplete and fragmented.

There has sometimes been a discrepancy between psycholinguists and computational linguists in their emphasis on one or other of these stages, although the field of word form detection in the speech stream has seen some intergrated research. For example, after Saffran et al.'s [7] striking results on the performance of young infants, the hypothesis that infants may invoke the *transitional probability* between phonemes to detect words has been supported by a number of subsequent experiments (e.g. [8], [9]). The core idea has subsequently been taken up by computational linguists, and a number of algorithms have been proposed, some of which show promising results (e.g. [10], [11]).

In the area of word-denotation association, while psychological studies abound (see e.g. [1], [12] for references), computational modelling has only started to take off relatively recently with more multi-modal data available, though largely restricted to the association of nouns uttered in isolation with 'Spelke objects' [13] —coherent middle-sized objects like toys and pieces of furniture. This state of affairs is not without reason,

given the problem of *referential uncertainty*: it is technically difficult to associate a word in a multi-word utterance with its intended referent. Computational models of identifying a word reference are proposed (e.g. [14], [15]), but a reliable platform on which to extract word-denotation associations in an interactive environment is yet to be established.

The situation is reversed, if again understandably, for grammar learning, which concerns primarily internal structures of linguistic expressions, the *compositionality* in particular, of compound expressions. Computational models have been proposed for unsupervised grammar induction, some logically oriented like ours (e.g. [16], [17]) although these presuppose fully word-segmented string. However there is a dearth of psycholinguistic studies, presumably due to the difficulty in extracting semantics of compound expressions from the empirical data in a meaningful manner.

The methodology of cognitive robotics could fill these gaps. Embodied robots with a speech module can offer a platform on which to test for plausible word-denotation association through a multitude of modalities. By exploiting such results, we can then test various algorithms of grammar induction dynamically on a robot under an interactive scenario. To our knowledge, while research in robotics on language learning that encompasses grammar is underway elsewhere [18]–[21], these studies focus on the learning architecture of an autonomous robot. Our research is unique in putting the robot in an interactive learning environment with human participants, acting as teachers, who are not imposed any predetermined restrictions for their speech. Our main focus is to create a model of natural human language acquisition and demonstrate the possible ways in which children learn grammar on the platform of cognitive robotics, rather than offering an engineered solution for an autonomous language learner, so that not only the field of robotics but also cognitive and social sciences can benefit. The objective of this paper is to propose and test such a model.

III. Implemented Model

A. Word sound form detection

Pre-linguistic children, presented with continuous stream of adult speech, are faced with the task of discovering the 'meaningless' sound units and patterns specific to the language they are exposed to, and extracting 'meaningful' sound forms that potentially correspond to words. This 'duality of patterning' is recognised as an overarching feature of human languages, and there is ample evidence that children attempt at the first, purely phonological part of the task prior to and without involving the latter, semantic part. Therefore, although we are aware that a number of higher-level factors could potentially contribute to word discovery, our model proposed here treats this first of the three stages as only involving the phonological recognition capacities *without* syntax or semantics. Specifically, we consider statistical grouping and prosody to play a central role.

The involvement of statistics in early word recognition is suggested in psycholinguistic evidence cited above. Hence our model incorporates statistically oriented procedures for word boundary detection ('word segmentation') in order to bootstrap a small lexicon of sound units. Amongst other alternative statistical measures with varying levels of sophistication, we employ a technnique developed by [10], which relies on relatively uncontroversial phonotactic and learning strategies, as our focus is empirical plausibility in word detection.

We nevertheless only employ such statistical word segmentation procedures as a secondary measure, because they presuppose all the constituent phonemes to be present beforehand. This appears implausible as a model of phonological acquisition as children are unlikely to possess the whole inventory of available phonemes at this stage. Instead, therefore, we invoke prosodic features for phonological bootstrapping, recognising the characteristics of the typical speech of teaching adults, Child-Directed Speech or CDS. That is, we take advantage of the 'affective prosody' typically found in CDS: a relatively long duration, high pitch contour or intensity on the stressed syllable of the word the teaching adult attempts to draw attention to. After selecting the syllable that receives these features, we then apply the word boundary detection procedure to 'construct' a word that contains that syllable, by finding the likely starting and ending points of the word.

In sum, our current model discovers 'word-like units' with a combination of prosodic and statistical approaches, i.e. by finding sequences of phonemes that are frequently coherent and possess a relatively longer duration and high and heavy intonation contour. There can surely be other contributory factors that could improve the performance of word detection, some of which are currently under investigation. One such reinforcing factor we consider is the feedback a carer provides to a prelinguistic, babbling infant, through which the infant is encouraged to reproduce the babbling that closely corresponds to a word. Another possibility we investigate is to chunk an utterance into intermediate *phonological phrases*[1] first before word segmentation, as a number of studies indicate children exploit them as a facilitating cue for word discovery [24].

B. Word-denotation association

CDS appears to be specifically tailored to the perceived level of linguistic skill of the child [25] in that adults speak more slowly and often repeat a portion of what has been said during the interaction. In such interactions, children learn new words with very few presentations. In order to achieve such fast learning, the infant's learning experience needs to be biased in some way. We therefore introduce *action* —speech in particular— on part of the learner: a factor that would be difficult to introduce without a cognitive robot. A perception-capable robot can decide what a sound unit that it considers a word may stand for and *try out* uttering it. Our hypothesis is that, following our previous work on social learning of behaviour acquisition [26], the adult utterances situated in the learning context, together with reinforcement via forms of affective feedback and communicative success or failure (e.g. in shared intentional reference), allow enough bias for fast learning to take place.

[1] A phonological phrase is a unit between a clause and a word in 'prosodic hierarchy' proposed by [22]. It can be realised and hence is delimitable phonetically in various ways, including lengthening of its last syllable [23].

In our model, learning relies on the association of words with sensorimotor and internal state information on one hand, and the learner's ability to share context via a form of rudimentary referential 'intent' with a human teacher on the other. These associations can then be scaffolded via regularities. More specifically under the HRI experiment scenario, the first step is to match the two modalities, i.e. what was said to the robot and what was experienced by the robot at that time. Whenever it goes through a similar experience, it utters the word most strongly associated with that experience. In our interactive teaching scenario, if the word uttered is appropriate or inappropriate, the teacher tends to 'encourage' the robot by repeating the word or 'correcting' it by replacing the word with an alternative, thereby reinforcing or weakening the association.

We are aware that there are a number of technical challenges to achieve such associations in the face of *referential uncertainty*, the issue of which word should correspond to which segment of experience. We make no pre-programmed choices as to which word or which segment of the sensorimotor stream is relevant for the robot, but simplify, to cope with these difficulties, the learning environment in the following three ways. First, we limit the 'target objects' to learn the words for, by using a limited set of objects (which in our experiment are, as we shall see, six shapes depicted on a cube, such as triangle and circle), which the robot has been pre-trained to categorise in advance. It nevertheless has *not* yet associated its sensorimotor attributes either with the shapes or any words at the start of learning and this association constitutes a new task.

Secondly, we take advantage of an information theoretic measure to deal with the issue of the choice of the relevant sensorimotor segment. Having associated a word with a sensorimotor stream, we then compute the 'information gain' [27] between the sensory attributes and the chosen word (a measure of 'mutual information' indicating the expected amount of information that effectively discriminates the given word by the given attribute). The values of this measure obtained in one session of learning is used during the following sessions to weight the similarity measure of current vs. stored sensorimotor experiences and generalises the experience.[2] Thus for example the word 'triangle' is associated with the triangle shape no matter where it is viewed and effectively de-weighting the head sensorimotor feedback in favour of the object identifier. In our study sensorimotor attributes which do not affect the primary association of object and word (such as the head proprioception and the coordinates of the object in the image) should thus become less relevant over time.

Thirdly regarding the issue of word choice in an utterance, we have exploited simple heuristics and choose to focus only on the last phonological phrase of an utterance and select the most heavily stressed word, usually the very last word. This may appear too limiting an approach, and more sophisticated, typically multi-modal algorithms (e.g. [14]) would eventually be required. However, we believe this decision is justifiable as the first approximation because it is observed in psycholinguistic literature that attention is drawn to new nouns by English-speaking adults by placing them at the end of an utterance [25, pages 62-66]. Our data, gathered from our HRI experiment sessions, also confirms this tendency, as we shall see in Section IV. It also fits the idea of 'incremental' word learning, in that children start out with a very small initial denotation-associated vocabulary and then build it up by adding other words employing a variety of methods. Thus our assumption amounts to maintaining that it is only at this 'bootstrapping' stage when the learner needs such a restricted focus. After acquiring a small set of words, they become able to recognise these words anywhere in an utterance, and expand their focus to other parts of it.

The net result of these procedures is that the robot human interactions are sufficiently biased to allow an association between the robot's own sensorimotor experience and the relevant word expressed by the human.

C. Logic-based grammar learning

In light of the observation that infants' early utterances could be interpreted to conform to a simple bipartite logical form ('pivot schema'), we have chosen to test the possibility that they also employ a similar strategy to build a grammar from an adult utterance: breaking it down into two. To model such a possiblity, we use a semantic representation with only two types, the object denotation and the predicate. We invoke a machinery from formal semantics [4] to represent these types: the *entity* (e) type and the function type that takes this type as an input to produce a truth value (*t*) as its output, or $e \rightarrow t$ type.

To model the 'incompleteness' of infants' initial grammar, we must abstract the standard logical representation and underspecify the predicate in terms of arity, i.e. how many arguments it may take. The number of entities involved in a logical formula, accordingly, is also unspecified. Thus an infant's incomplete understanding of an utterance is modelled as some relation (or property) holding of one or more objects present in the learning situation. A Kleene star notation is used to denote this underspecified variety of predicate: $(e \rightarrow)^*t$ (for reason of brevity, an abbreviated notation of e^*t will be used henceforth). Similarly, the unspecified number of arguments will be notated as e^*.

We define the learner's task as partitioning the recognised words in an utterance into two disjoint and exhaustive subsets and associating each subset with one of the two types. For example, if the utterance is the sequence $w_1, ..., w_n$ and the detected words are w_2, w_4 and w_5, the learner may partition them into $\{w_4\}$ and $\{w_2, w_5\}$, and pair up each of these sets with each of the two types, for example $\{w_4\}$ with e^*t and $\{w_2, w_5\}$ with e^*. For terminological convenience we call the combination of such two subsets a (two-way) *partition* of words.[3] Notice that it is a set of words, rather than words themselves, that are associated with types. The underlying assumption here is that each content word in an *utterance*

[2]For the experiments reported here, the sensorimotor stream consists of object location in Cartesian space, position reading of the robot's head pan, tilt and roll actuators, binary face detection identifier and the object identifier.

[3]As we shall see, it does not precisely correspond to the set-theoretic notion of partition, since we include an empty subset.

as a whole —that is including the unrecognised portion— is *part of* (albeit potentially the whole of) either an entity or a predicate of the learning target, or the 'correct' partition. For example, for the utterance '*Look at that red box on the table, John*', the content words that would be part of a noun phrase (*that red box*, *the table* and *John*), namely {*red*, *box*, *table*, *John*}, will be considered the target entity word set (e^*), while {*look*} the predicate word set (e^*t). Towards this target partition, the learner's partition will 'grow' with the growth of its lexicon, and hence of the recognised portion.

As an example, let us suppose that the learner has been exposed to this example utterance ('*Look at the red box on the table, John'*) and recognises the following words, *look*, *box* and *John*. As the learner's word recognition is only partial, the utterance can be represented as a mixed list of phonemes (unrecognised portion) and words (recognised portion, with orthographs in square bracket) as below:

[*look*], ətðərɛd, [*box*], ɔnðətɛɪbl, [*John*]

The possible partitions of the recognised words are as follows:

{*look, box, John*}, {}
{*look, box*}, {*John*}
{*look, John*}, {*box*}
{*look*}, {*box, John*}

Then in our association task, each subset in a partition is paired up with either of the two types, which can potentially produce $2^3 = 8$ combinations:

1) e^*t: {*look, box, John*}, e^*: {}
2) e^*: {*look, box, John*}, e^*t: {}
3) e^*t: {*look, box*}, e^*: {*John*}
4) e^*: {*look, box*}, e^*t: {*John*}
5) e^*t: {*look, John*}, e^*: {*box*}
6) e^*: {*look, John*}, e^*t: {*box*}
7) e^*t: {*look*}, e^*: {*box, John*}
8) e^*: {*look*}, e^*t: {*box, John*}

The learning agent then chooses one of the options, and sets out to induce the likely grammar. Postponing at present the discussion on how to narrow down these potentially numerous options, we first describe how grammar can be gleaned out of one selected partition. Another necessary ingredient is some representation of meaning hypothesis: $P(c_1, ..., c_n)$ being true, where P stands for a hypothesised predicate (whose exact content and arity are yet to be established) and $c_1, ..., c_n$ for an unspecified number of constants (whose references also may not yet have been established). Informally, it only says some relation holds of some objects. Suppose now that Option 6 is selected from the above example, where the learner supposes 'box' to denote the relation P and 'look' and 'John' to denote objects, say c_1 and c_3. Then the learner has effectively built a hypothesis on grammar in two pieces: (1) *look* and *John* together refer to at most two entities and (2) combine them with *box* and one will get a sentence,[4] except that, unfortunately, it is wrong ('box' is not a predicate that can take two arguments).

[4]More precisely, 'box' is such a word that if you combine it with at least two entity-type words, one will get a sentence.

In this model, in which the learner learns incrementally, s/he does not *have to* know what exactly the denotations of words are in order to induce the semantic types, but it *helps* to know in advance what the denotation of a word or some words may roughly be, in order to restrict his/her hypothesis space and exclude the wrong ones. As we saw, there can be many options for the word-set/type combination task (2^n combinations for n words), which may well require an untenable memory load. We need some bias(es) to make the learner's hypotheses converge. Now if, for example, the learner is confident enough to think that *box* denotes some object, then some wrong options, including our example above, are eliminated and the hypothesis space is reduced significantly (by half, with Options 1, 3, 6 and 8 eliminated). This corresponds to what Pinker [3] called the effect of 'semantic bootstrapping'.

We are in a vantage position to be able to exploit some object-word associations independently obtained in a manner described in preceding subsection III-B, i.e. by correlating an embodied robot's internal sensorimotor states with some pre-defined objects. This generates the semantic bootstrapping effect, reducing the number of hypotheses. Therefore, we primarily test whether and to what extent the right type-word associations can be made with this bias. The following section summarises the experimental method and results.

IV. Experiment and evaluation

A. Experiment

We conducted a series of one-to-one HRI experiments with the Kaspar2 robot [28], a humanoid robot approximately of the size of an infant. We had a total of eight participants (5 female, 3 male) and each was asked to verbally 'teach' it about six shapes drawn on the six planes of a cube *as if* the robot were a 1-2 year old child (see figure 1) for two minutes in one session. The shapes are circle, heart, moon, square, sun and triangle. No other instruction was given concerning the way s/he should speak to the robot. Each participant went through a total of five of such sessions.

The robot was pre-programmed to track and habituate for a given period on these shapes. In the first session for each participant the robot only 'listens' to what s/he says, though it responds by moving its head if s/he moves the cube, as it already recognises these shapes. Between the sessions, the speech data was put through a speech recogniser and analysed in a manner to be described in the following section. Thus the learning procedure was applied off-line and robot learning proceeded separately for each participant. It 'starts talking' from the second session onwards by making 'one-word utterances' according to the method described in Section III-B. This usually elicits the response from the participant, making the sessions more 'interactive'.

B. Data processing and analysis

Following each session the speech stream of the human was converted into phonemes, and the methods described in III-A were applied to find words from these phonemes. Then the procedures described in III-B selected the utterance-final stressed words and aligned them with the sensorimotor

Fig. 1. ***Sharing Reference with a Teacher in Context.*** *Kaspar2 interacts with the teacher, where both the robot and the human share attention on the shapes on the box, a case of rudimentary shared intentional reference.*

modalities experienced by the robot during the interaction session.

This processed modality stream became the basis for the robot to use its learnt experiences for the subsequent sessions. From Session 2 onwards the robot was allowed to match its current sensorimotor input against that learnt in the previous sessions and react to the human by uttering a word at a time. For example, when presented with the moon shape, if the human tutor had previously described it as 'moon', the robot might say 'moon'.

We started to analyse the data for grammar acquisition —semantic type mapping described in III-C— after all the experiment sessions. Thus the learning took place off-line. The learning task is to output a grammar hypothesis, i.e. a partition (two subsets) of the recognised words in an utterance and its association with the two types. Each word in the subset associated with a type, either e^*or e^*t, is said to belong to, or part of, that type. We require the learner to produce a single hypothesis. Therefore, we give some biases (to be described in the next subsection) to the learner, and let the learner randomly choose from the remaining hypotheses.

C. Biases in grammar learning

As discussed in Sections III-B and III-C, we pre-allocated the *e* type to the words that have been aligned with our pre-defined six objects, i.e. put them in the e^* subset of the two-way partition. This does not mean, however, that there are six such words, because the association is made only indirectly through sensorimotor stream. Any utterance-final word with a strong enough association with one of the six shapes may be allocated, potentially erroneously, with the *e* type. This mapping of some words with the *e* type constitutes the semantic bootstrapping effect.

To see whether and to what degree they help the hypotheses to converge further, we have tested two further biases. One concerns the case in which there is only one word in an utterance (and hence is not about 'combinations' in an intuitive sense): that the learner prefers this single word to correspond to *e* type. We call this 'Object Bias'. The other is to prefer the subset with fewer words to correspond to e^*t and the larger subset to e^*. The rationale behind this is that it would be more likely that the learning child initially takes the utterance to be a simple main clause, rather than considering the possibility of a complex predicate (e.g. 'x runs or walks') or an embedded predicate (e.g. 'x thinks that y runs'). We call it 'e^*t<e^* Bias'.

D. Results and discussion

1) Data overview: We have compiled 40 transcripts (of phonemes, 8 particinpants $\times$5 sessions) of two-minute sessions. In total there were 8142 word token occurrences and 1671 utterances, averaging 203.55 words and 41.78 utterances in each session, giving the average length of an utterance at 4.87 words. The number of word type lemmas (i.e. in terms of the 'words' in dictionary entries) of these 8142 occurrences is 3215. The data can thus be characterised as a collection of short utterances with many repetitions of the same word type lemmas.

In relation to utterance-final words discussed in III-B, indeed none of the words that refer to the six key objects failed to appear utterance-finally. In fact 43% of all the token occurrences of these words appeared utterance-finally. We should also add that this is partly due to the fact that a large proportion of utterances were elliptical in our data, another hallmark of CDS. 6.8% of the utterances were a bare NP (noun phrase consisting of a single noun or pronoun alone) or an NP consisting of a determiner and a single noun.

2) Detected words and 'semantic bootstrapping': Below are sample lists (for two participants) of detected words from the phoneme strings. The numbers on the 'Session' column indicate that the listed words have been learnt *between* these two sessions. We only show those words are 'newly learnt' in addition to those already learnt up to the preceding session. On average, 31.23 word types have been 'detected' after five sessions for one participant.

Session	Participant 1	Participant 2
1-2	and, box*, black, dot, this, four, got, heart*, wait, again, sun, circle*, triangle*, get, three, Kaspar, you	called, circle*, circles*, done, its, Kaspar, look, lots, moon*, of, one, see, shape*, size, square*, star*, that*, triangles*
2-3	big, it, triangle*, triangles, see, shapes*, middle	angle, bits, bumps, this, got, heart*, has, you, sky, square*, sides
3-4	bottom, edge, five, have, corners, points, taken, square*, top	can't, now, right, round*, sun*, that, two
4-5	back, bumps, when, inside, is, can, like, say	about, crescent*, minutes, left, okay, rainbow*

Fig. 2. Phonologically detected and semantically bootstrapped (*) words

The data shown above demonstrate that the words for our key six objects, usually nouns, are amongst the first ones to be detected. The fact that these words were often the most prosodically prominent in an utterance greatly contributed to their early detection. Related to this is the fact mentioned earlier that these words tended to occur in elliptical one-noun fragments, i.e. bare NPs or NPs with a deterimner and a noun. In the former case that sole occurrence of a noun automatically counts as the most prosodically prominent. In the latter, the unaccented determiner is usually stripped off, leaving the noun to be detected.[5]

[5]More precisely this is a combined effect of the statistical and prosodic procedures described earlier. The statistical procedure splits the two-word NP into a determiner and a noun, while the relative prosodic prominence picks up the noun. This is of course except the case that the determiner itself is accented. Such an accent tends to fall on the demonstrative or contrastive use of a determiner. This may have contributed to the fact that *this* and *that* have been detected early in our data.

The words marked with * indicate 'semantically bootstrapped' ones, i.e. ones associated with one of the six objects strong enough for the robot to say those words, and hence placed in the e^* subset. On average there were 7.6 such words (lemmas) for one participant. They tended to have been learnt relatively quickly, usually by the second session. The frequency of these words were found to be very high: they account for 7.85% against the total number of word token occurrences.

The threshold for the association 'strength' has been set so as not to make the robot too talkative or too reticent (about one single-word utterance at each 'event' like a hand movement of the participant), so there is a certain element of arbitrariness in cognitive terms. However, the words over this threshold do coincide closely with the common nouns associated with the objects, suggesting that the relative strengths of sensorimotor streams are good indicators of object salience.

Furthermore, as predicted, the robot's performance on 'naming the right object' improved over time, confirming the reinforcement effect. However, it still needs to be ascertained whether this reinforcement is mainly due to sensorimotor information, or to interaction such as 'corrections' by the tutor.

3) Type mapping learnt: Below are the F-scores of the word-type mapping in comparison with the manually annotated gold standard of Part-of-Speech tagging. Here the precision is the proportion of the correctly classified words overall, while the recall is that of the words correctly classified into a cateogry (e^*or e^*t) in that category. The semantically bootstrapped words that were pre-allocated a type beforehand are excluded when calculating the scores. We considered common/proper nouns, determiners and attributive adjectives[6] to belong to the portion of the e^* type, and all the other parts of speech to that of the e^*t type.

	F-Score (P:precision, R:recall)
Baseline (no bias)	$.4675(P:.4502, R:.4863)$
Semantic bootstrapping	$.6019(P:.5801, R:.6254)$
+ Object Bias	$.6417(P:.6190, R:.6663)$
+ $e^*t \leq e^*$ Bias	$.6987(P:.6753, R:.7239)$

The baseline score confirms the randomness of the mapping task. There is an improvement of 23.44 percentage points with semantic bootstrapping, indicating that pre-allocating a type to a handful of words can greatly improve the performance. Admittedly this result was obtained with a distributional skew and repetitions in the data as described earlier and should not be hastily generalised, but these aspects represent typical characteristics of CDS. There were further improvements for both additional biases we have tested, Object Bias to a lesser degree than $e^*t \leq e^*$ Bias, suggesting the relevance of such biases.

V. Concluding remarks and future tasks

We have proposed the three-stage model for language acquisition that differentiates three aspects of learning: phonological, word-object association and grammar induction. We have also described the method of cognitive robotics we employed and reported on the initial results of our HRI learning experiment.

We are still at an early stage of an ongoing research programme of a multi-disciplinary nature that requires refinements and further incorporation of research findings in a number of areas. We therefore conclude the present paper with some remarks on future tasks, with an emphasis on interactivity, a unique feature of our model.

One purely technical, if important, restriction of the experiment thus far is that we have been applying both lexical meaning and grammar learning procedures off-line. Practical problems such as cross-language interface have prevented us from on-line experimentation, but we plan to conduct further experiments in a fully on-line manner as these problems are overcome.

Even the area of word sound form association still leaves much room for interaction to play a role. For example, Oudeyer [29] points out the importance of infant babbling as a stage in language acquisition, and the development of shared speech codes. From this standpoint we are simulating the interaction between a babbling infant and a carer, with a 'reward structure' in place when the child produces a word by chance.

For word-denotation association, we plan to introduce the mechanism of reinforcing or weakening the learner's hypotheses through 'joint attention' shared by the teacher, which has long been observed to be implicated in early language learning [30]. More specifically, the use of an eye tracker to identify the human interactor's gaze direction should be useful for disambiguating word-object associations in interactive scenarios involving multiple objects.

Regarding grammar induction, the grammars learnable by the presented method remain a rudimentary, bipartite 'pivot schema' variety, and hence need to be expanded at the very least to capture multiple-argument structures, recursion and word order. Further, our model still lacks the most basic grammatical features carried by function words, e.g. determiners and quantifiers. However, we envisage the semantics-based learning to be able to capture such basic features by gradually introducing more sophisticated machineries of formal semantics. One potentially promising direction is a shift towards property-based learning. If we could treat objects as 'bearers' of properties, the model could handle various phrases in a more unified and consistent manner (e.g. predicative adjectives as fresh property assignment to an object and noun phrases as generalised quantifiers). Moreover, such a model would be more sensitive to evaluative and corrective feedbacks ('Well done', 'No, it's y' (instead of x), etc.), as the learner could then enhance or weaken his/her hypotheses by correlating this apparent communicative success/failure with his/her phenomenal experience.

[6]Adjectives are thus treated differently depending on whether they are attributive or predicative. See the last section for brief discussion.

Acknowledgement

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

REFERENCES

[1] M. Tomasello, *Constructing a Language*. Harvard UP, 2003.

[2] R. Kempson, W. Meyer-Viol, and D. Gabbay, *Dynamic Syntax*. Blackwell, 2001.

[3] S. Pinker, *Language Learnability and Language Development*. Harvard University Press, 1984.

[4] R. Monatague, *Formal Philosophy*, R. Thomason, Ed. Yale University Press, 1974.

[5] D. Swingley, "Contributions of infant word learning to language development," *Philosophical Transactions of the Royal Society*, vol. 364, pp. 3617–3632, 2009.

[6] M. Braine, "The ontogeny of English phrase structure," *Language*, vol. 39, pp. 1–14, 1963.

[7] J. Saffran, E. Aslin, and R. Newport, "Word segmentation: The role of distributional cues," *Journal of Memory and Language*, vol. 35, 1996.

[8] E. D. Thiessen and J. R. Saffran, "When cues collide: use of stress and statistical cues to word boundaries by 7- to 10-month-old infants," *Developmental Psychology*, vol. 39, pp. 706–716, 2003.

[9] D. Swingley, "Statistical clustering and the contents of the infant vocabulary," *Cognitive Psychology*, vol. 50, pp. 86–132, 2005.

[10] M. Brent and T. Cartwright, "Distributional regularity and phonotactic constraints are useful for segmentation," *Cognition*, vol. 61, pp. 93–125, 1996.

[11] S. Goldwater, T. Griffiths, and M. Johnson, "A Bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, pp. 21–54, 2009.

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

[13] E. Spelke, "Initial knowledge: six suggestions," *Cognition*, vol. 50, pp. 443–447, 1994.

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

[15] N. Chang, "Constructing grammar: A computational model of the emergence of early constructions," Ph.D. dissertation, University of California at Berkeley, 2008.

[16] P. W. Adriaans and M. Vervoort, "The EMILE 4.1 Grammar Induction Toolbox," pp. 293–295, 2002.

[17] S. A. Fulop, "Semantic bootstrapping of type-logical grammar," *Journal of Logic, Language and Information*, vol. 14, pp. 49–86, 2004.

[18] M. McClain, "Semantic based learning of syntax in an autonomous robot," *International Journal of Humanoid Robotics*, vol. 4, pp. 321–346, 2007.

[19] Y. Sugita and J. Tani, "Learning semantic combinatoriality from the interaction between linguistic and behavioral processes," *Adaptive Behavior*, vol. 13, no. 1, pp. 33–52, 2005.

[20] L. Steels, "The emergence and evolution of linguistic structure: from lexical to grammatical communication systems," *Connection Science*, vol. 17, no. 3-4, pp. 213–230, 2005.

[21] P. F. Dominey, "Emergence of grammatical constructions: evidence from simulation and grounded agent experiments," *Connection Science*, vol. 17(3-4), pp. 289–306, 2005.

[22] E. Selkirk, *Phonology and Syntax*. MIT Press, 1984.

[23] C. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P. J. Price, "Segmental durations in the vicinity of prosodic phrase boundaries," *Journal of the Acoustical Society of America*, vol. 91, pp. 1707–1717, 1992.

[24] A. Gout, A. Christophe, and J. Morgan, "Phonological phrase boundaries constrain lexical access ii: Infant data," *Journal of Memory and Language*, vol. 51, pp. 548–567, 2004.

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

[26] J. Saunders, C. L. Nehaniv, K. Dautenhahn, and A. Alissandrakis, "Self-imitation and environmental scaffolding for robot teaching," *International Journal of Advanced Robotics Systems*, vol. 4(1), pp. 109–124, 2007.

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

[28] 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*, vol. 6(3,4), pp. 369 – 397, 2009.

[29] P.-Y. Oudeyer, *Self-Organization in the Evolution of Speech*. OUP, 2006.

[30] M. Tomasello and M. J. Farrar, "Joint attention and early language," *Child Development*, vol. 57, no. 6, pp. 1454–1463, 1986.