ARTICLE IN PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Cognitive Systems Research xxx (2004) xxx–xxx

Cognitive Systems RESEARCH

www.elsevier.com/locate/cogsys

Developmental stages of perception and language acquisition in a perceptually grounded robot

Action editor: Luc Berthouze

Peter Ford Dominey *, Jean-David Boucher

Institut des Sciences Cognitives, CNRS UMR 5015, 67 Boulevard Pinel, 69675 Bron Cedex, France

Received 1 July 2004; accepted 6 November 2004

Abstract

The objective of this research is to develop a system for language learning based on a "minimum" of pre-wired language-specific functionality, that is compatible with observations of perceptual and language capabilities in the human developmental trajectory. In the proposed system, meaning (in terms of descriptions of events and spatial relations) is extracted from video images based on detection of position, motion, physical contact and their parameters. Meaning extraction requires attentional mechanisms that are implemented from low-level perceptual primitives. Mapping of sentence form to meaning is performed by learning grammatical constructions, i.e., sentence to meaning mappings as defined by Goldberg [Goldberg, A. (1995). *Constructions*. Chicago and London: Univ. of Chicago Press]. These are stored and retrieved from a "construction inventory" based on the constellation of grammatical function words uniquely identifying the target sentence structure. The resulting system displays robust acquisition behavior that reproduces certain observations from developmental studies, with very modest "innate" language specificity.

Keywords: Language acquisition; Event perception; Grammatical construction; Neural network

1. Introduction

A challenge of epigenetic robotics is to demonstrate the successive emergence of behaviors in a developmental progression of increasing processing power and complexity. A particularly interesting avenue for this methodology is in language processing. Generative linguists have posed the significant challenge to such approaches via the claim that the learning problem is too underconstrained and must thus be addressed by a highly pre-specified Universal Grammar (Chomsky, 1995). The current research proposes an alterna-

* Corresponding author. Tel.: +33 437 911212; fax: +33 437 911210.

E-mail addresses: dominey@isc.cnrs.fr (P.F. Dominey), boucher@isc.cnrs.fr (J.-D. Boucher).

URL: http://www.isc.cnrs.fr/dom/dommenu-en.htm.

1389-0417/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.cogsys.2004.11.005

tive, identifying a restricted set of functional requirements for language acquisition, and then demonstrating a possible framework for the successive emergence of these behaviors in developmentally plausible systems, culminating in a grounded robotic system that can learn a small language about visual scenes that it observes.

1.1. Functional requirements

We adopt a construction-based approach to language in which acquisition is based on learning mappings between grammatical structure and meaning structure (Goldberg, 1995). In this context, the system should be capable of: (1) extracting meaning from the environment, (2) learning mappings between grammatical structure and meaning, and (3) identifying-discriminating between different grammatical structures of input sentences. In the following sections, we outline how these requirements can be satisfied in a biologically and developmentally plausible manner.

In this developmental context, Mandler (1999) suggested that the infant begins to construct meaning from the scene based on the extraction of perceptual primitives. From simple representations such as contact, support, and attachment (Talmy, 1988), the infant could construct progressively more elaborate representations of visuospatial meaning. In this context, the physical event "collision" can be derived from the perceptual primitive "contact". Kotovsky and Baillargeon (1998) observed that at 6 months, infants demonstrate sensitivity to the parameters of objects involved in a collision, and the resulting effect on the collision, suggesting indeed that infants can represent contact as an event predicate with agent and patient arguments. Similarly, Quinn, Polly, Furer, Dobson, and Nanter (2002) have demonstrated that at 6–7 months, infants are sensitive to binary spatial relations such as above and below.

Bringing this type of perception into the robotic domain, Siskind (2001) has demonstrated that force dynamic primitives of contact, support, and attachment can be extracted from video event sequences and used to recognize events including pick-up, put-down, and stack based on their characterization in an event logic. Related results have been achieved by Steels and Baillie (2002). The use of these intermediate representations renders the systems robust to variability in motion and view parameters. Most importantly, this research demonstrated that the lexical semantics for a number of verbs could be established by automatic image processing.

Once meaning is extracted from the scene, the significant problem of mapping sentences to meanings remains. The nativist perspective on this problem holds that the ⟨sentence, meaning⟩ data to which the child is exposed is highly indeterminate, and underspecifies the mapping to be learned. This "poverty of the stimulus" is a central argument for the existence of a genetically specified universal grammar, such that language acquisition consists of configuring the UG for the appropriate target language (Chomsky, 1995). In this framework, once a given parameter is set, its use should apply to new constructions in a generalized, generative manner.

An alternative functionalist perspective holds that learning plays a much more central role in language acquisition. The infant develops an inventory of grammatical constructions as mappings from form to meaning (Goldberg, 1995). Developing in the second year of life, these constructions are initially rather fixed and specific, and later become generalized into a more abstract compositional form employed by the adult (Tomasello, 1999, 2003). In this context, construction of the relation between perceptual and cognitive representations and grammatical form plays a central role in learning language (e.g., Feldman, Lakoff, Stolcke, & Weber, 1990; Feldman et al., 1996; Langacker, 1991; Mandler, 1999; Talmy, 1988; Tomasello, 1999, 2003).

These issues of learnability and innateness have provided a rich motivation for simulation studies that have taken a number of different forms. Elman (1990) demonstrated that recurrent networks are sensitive to predictable structure in grammatical sequences. Subsequent studies of grammar induction demonstrate how syntactic structure can be recovered from sentences (e.g., Stolcke & Omohundro, 1994). From the "grounding of language in meaning" perspective (e.g., Feldman et al., 1990, 1996; Goldberg, 1995; Langacker,

1991) in which language is characterized in terms of its communicative function, rather than in purely formal terms, Chang and Maia (2001) exploited the relations between action representation and simple verb frames in a construction grammar approach, and Cottrel, Bartell, and Haupt (1990) associated sequences of words with simple image sequences. In an effort to consider more complex grammatical forms, Miikkulainen (1996) demonstrated a system that learned the mapping between relative phrase constructions (e.g., "The dog that the cat chased bit the kid") and the corresponding multiple event representations, based on the use of a stack for maintaining state information during the processing of the next embedded clause in a recursive manner.

In a more generalized approach, Dominey (2000) exploited the regularity that sentence to meaning mapping is encoded in all languages by word order and grammatical marking (bound or free) (Bates, McNew, MacWhinney, Devescovi, & Smith, 1982). The model is based on the functional neurophysiology of cognitive sequence and language processing and an associated neural network model that has been demonstrated to simulate interesting aspects of infant (Dominey & Ramus, 2000) and adult language processing (Dominey, Hoen, Lelekov, & Blanc, 2003). As will be described below, the model formalizes generalized mappings from sentence structure to meaning structure (grammatical constructions – Goldberg, 1995), and indexes these constructions based on the configurations of closed class words in the different sentence types. The model in the current study extends that work in the context of epigenetic development to address complex hierarchical structure, and spatial relations in a grounded learning environment.

1.2. Objectives

The goals of the current study are threefold: First to test the hypothesis that meaning – in terms of predicate-argument descriptions of events and spatial relations – can be extracted from visual scenes based on the detection of contact and its parameters in an approach similar to but significantly simplified from Siskind (2001); Second to use these meanings in ⟨sentence, meaning⟩ pairs as inputs to the grammatical construction learning model of Dominey (2000) in order to demonstrate that these two systems can be combined to perform miniature language acquisition; and finally to demonstrate that the combined system can provide insight into the developmental progression in human language acquisition without the necessity of a pre-wired parameterized grammar system (Chomsky, 1995).

1.3. The behavioral learning context

As illustrated in Fig. 1, the human experimenter enacts and simultaneously narrates visual scenes made up of events that occur between a red cylinder, a green block and a blue semicircle or "moon" on a black matte table surface. A video camera above the surface provides a video image that is processed by a color-based recognition and tracking system (Smart-Panlab, Barcelona Spain) that generates a time ordered sequence of the contacts that occur between objects that is subsequently processed for event analysis (below). The simultaneous narration of the ongoing events is processed by a commercial speech-to-text (STT) system (IBM ViaVoice™). Speech and vision data were acquired and then processed off-line yielding a data set of matched sentence – scene pairs that were provided as input to the structure mapping model. For example, in Fig. 1, the human takes the blue moon and uses it to "give" the block to the cylinder. At the same time, the human narrates this, saying "The moon gave the block to the cylinder". The STT system yields the text string "The moon gave the block to the cylinder", and the event analysis produces a description of the form *gave*(*moon*, *block*, *cylinder*), thus yielding one ⟨sentence, meaning⟩ pair. A total of 300 ⟨sentence, meaning⟩ pairs were tested in the following experiments. The desired performance of the model is the ability to understand new sentences, i.e., when given a new sentence to accurately generate the corresponding meaning.

1.4. Roadmap

In the next sections, we will identify the functional requirements for the system, i.e., what it is

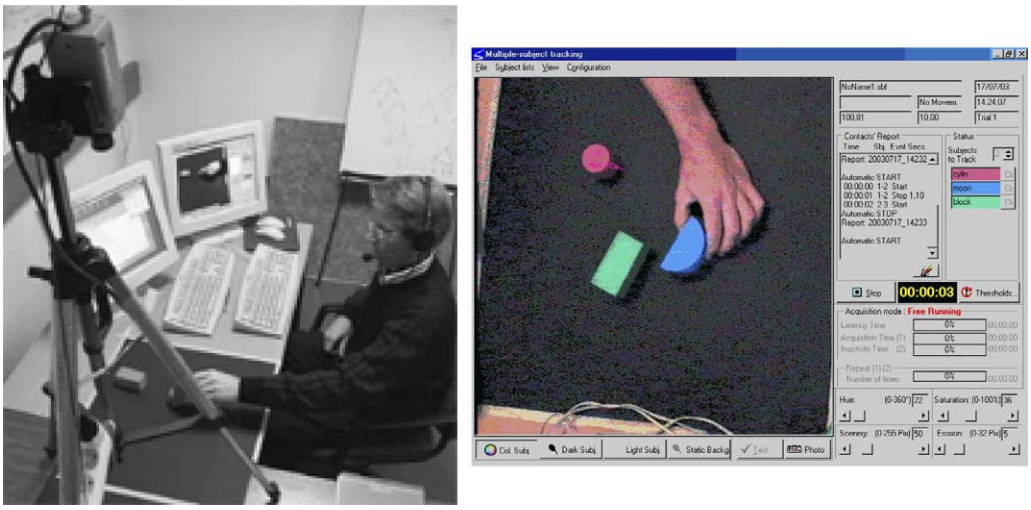

Fig. 1. Perceptually grounded robotic system, and view of the scene from the system's perspective.

to do. The first requirement will be to extract meaning from the environment, the second will be to map sentence structure onto this meaning for a given sentence type, and the third will be to generalize this to a variety of different sentence types or grammatical constructions. Experimental results with the system will then be presented, with extensions to complex grammatical constructions, and demonstration that the semantic predicate-argument encoding for events can naturally extend to allow the system to learn about spatial relations. The possible extension of these semantic representations to allow the emergence of new semantic structures is then discussed with respect to the spatial relation "between," followed by the general discussion.

2. Requirement 1: extracting meaning

For a given video sequence (see snapshot in Fig. 1), the visual scene analysis generates the corresponding event description in the format *event* (*agent*, *object*, *recipient*). The temporal schemas for the different events are displayed in Fig. 2.

2.1. Single event labeling

Events are defined in terms of contacts between elements. A contact is defined in terms of the time at which it occurred, the agent, object, and duration of the contact. The agent is determined as the causal element that had a larger relative velocity towards the other element involved in the contact, i.e., "the one that was moving towards the other one" as in the collision events of Kotovsky and Baillargeon (1998). Based on these parameters of contact, scene events are recognized as follows:

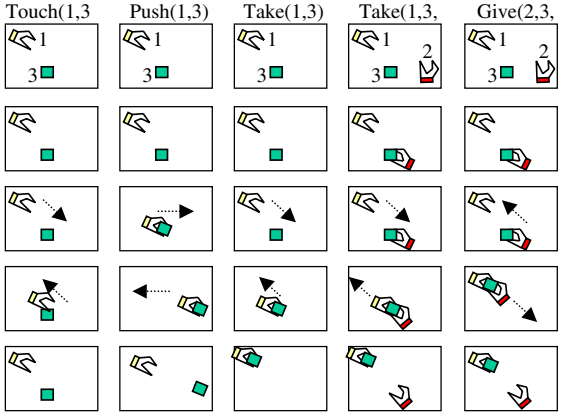

Fig. 2. Perceptual scene analysis for touch, push, take (with two and three arguments), and give. Each of these events is decomposed into a sequence of simple contacts. Agency is determined as the object that had a greater relative velocity in the contact.

Touch(*agent*, *object*): A single contact, in which (a) the duration of the contact is inferior to *touch_duration* (1.5 s), and (b) the *object* is not displaced during the duration of the contact.

Push(*agent*, *object*): Similar to touch, with a greater contact duration, superior or equal to *touch_duration* and inferior to *take_duration* (5 s), and object displacement.

Take(*agent*, *object*): A single contact in which (a) the duration of contact is superior or equal to *take_duration*, (b) the object is displaced during the contact, and (c) the agent and object remain in contact.

Take(*agent*, *object*, *source*): Multiple contacts, as the agent takes the object from the source. Same as Take(agent, object), and for the optional second contact between agent and source (a) the duration of the contact is inferior to *take_duration* and (b) the agent and source do not remain in contact. Finally, contact between the object and source is broken during the event.

Give(*agent*, *object*, *recipient*): Multiple contacts as agent takes object, then initiates contact between object and recipient.

These event labeling templates (illustrated in Fig. 2) form the basis for a template matching algorithm that labels events based on the contact list, similar to the spanning interval and event logic of Siskind (2001).

2.2. Complex "hierarchical" events

The events described above are simple in the sense that they have no hierarchical structure. This imposes serious limitations on the syntactic complexity of the corresponding sentences (Feldman et al., 1996; Miikkulainen, 1996). The sentence "The block that pushed the moon was touched by the triangle" illustrates a complex event that exemplifies this issue. We address this issue by explicitly coding such complex compound events as the set of their constituent events. Thus, the corresponding compound event will be recognized and represented as a pair of temporally successive simple event descriptions, in this case: *push*(*block*, *moon*), and *touch*(*triangle*, *block*). The "block" serves as the link that connects these two simple events in order to form a complex hierarchical event.

3. Requirement 2: mapping sentences to meaning

Our approach is based on the cross-linguistic observation that open class words (e.g., nouns, verbs, adjectives, and adverbs) are assigned to their thematic roles based on word order and/or the pattern of closed class words (grammatical function words or morphemes including prepositions and determiners) in the sentence (Bates et al., 1982).

The mapping of sentence form onto meaning for sentence comprehension takes place at two distinct levels: Words are associated with individual components of event descriptions and grammatical structure is associated with functional roles within scene events (Fig. 3). The first level has been addressed by Siskind (1996), Roy and Pentland (2002), and Steels (2001) and we treat it here in a relatively simple but effective manner. Our principle interest lies more in the second level of mapping between scene and sentence structure, and the ability to handle a large variety of different mappings, or grammatical constructions. Fig. 3(a) and (b) illustrates how two different grammatical constructions are processed by the model. The passive construction "*object* was *verb* to *recipient* by *agent*" in (a) and the active construction "*agent verb object* to *recipient*" in (b) both map (with different transformations) to the semantic representation of the event ACTION(AGENT, OBJECT, RECIPIENT) as illustrated. Eqs. (1)–(7) implement the model depicted in Fig. 3, and are derived from a neurophysiologically motivated model of sensorimotor sequence learning (Dominey, 2000; Dominey et al., 2003; Dominey & Hoen, 2005). In these equations, "=" designates an update of the left side by the right side. The associative memories are implemented as neural networks that correspond to modifiable cortico-cortico and cortico-striatal synapses. The ConstructionIndex corresponds functionally to a recurrent cortico-cortical network that has here been simplified for computational complexity reduction (see Dominey et al., 2003 for more extensive presentation of the underlying neurophysiology). Corresponding human neurophysiology can be seen in Hoen, Pachot-Clouard, Segebarth, and Dominey (2005) and Dominey and Hoen (2005). Once the model

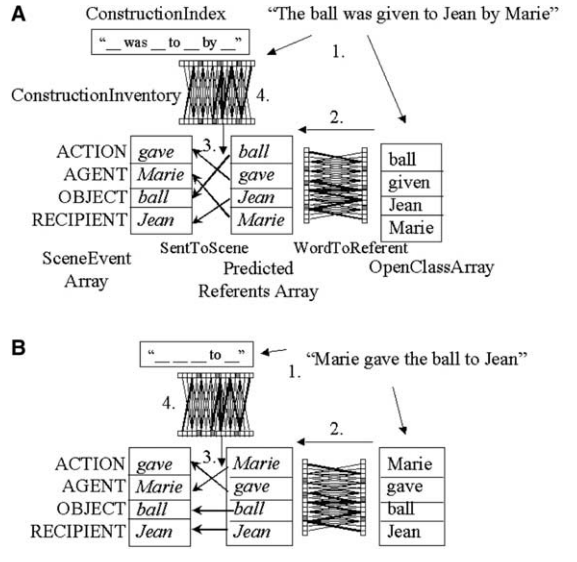

Fig. 3. Model Overview: Processing of active and passive sentence types in A, B, respectively. 1. On input, open and closed class words are segregated. Open class words populate the Open Class Array (OCA), while closed class words populate the ConstructionIndex. Visual Scene Analysis populates the Scene Event Array (SEA) with the extracted meaning as scene elements. 2. Words in OCA are translated to Predicted Referents via the WordToReferent mapping to populate the Predicted Referents Array (PRA). 3. PRA elements are mapped onto their roles in the Scene Event Array (SEA) by the SentenceToScene mapping, specific to each sentence type. 4. This mapping is retrieved from Construction Inventory, via the ConstructionIndex that encodes the closed class words that characterize each sentence type. Words in sentences, and elements in the scene are coded as single ON bits in respective 25-element vectors. Note the different SentToScene mapping for active and passive in A and B.

has been trained on well formed ⟨sentence, meaning⟩ pairs, it can then process new sentences that were not used in training (with the leaned vocabulary or lexicon) and generate for these sentences their corresponding meaning. This is the desired output processing of the trained model. Performance is measured by comparing this predicted meaning to the actual meaning that is provided in the ⟨sentence, meaning⟩ input pair.

3.1. Word meaning

Eq. (1) describes the associative memory, WordToReferent, that links word vectors in the OpenClassArray (OCA) with their referent vectors in the SceneEventArray (SEA). For all k, m, $1 \leqslant k \leqslant 6$, corresponding to the maximum number of words in the OCA, and $1 \leqslant m \leqslant 6$, corresponding to the maximum number of elements in the SEA. For all i and j, $1 \leqslant i,j \leqslant 25$, corresponding to the word and scene item vector sizes, respectively. In the initial learning phases, there is no influence of syntactic knowledge and the word-referent associations are stored in the WordToReferent matrix (Eq. (1)) by associating every word with every referent in the current scene ($\alpha = 1$), exploiting the cross-situational regularity (Siskind, 1996) that a given word will have a higher coincidence with referent to which it refers than with other referents. This initial word learning contributes to learning the mapping between sentence and scene structure (Eqs. (4)–(6)). Then, knowledge of the syntactic structure, encoded in SentenceToScene can be used to identify the appropriate referent (in the SEA) for a given word (in the OCA), corresponding to a zero value of α in Eq. (1). In the current studies, this transition is made manually. In actual development, a threshold of confidence in the syntactic knowledge could be used to determine this transition automatically. In this "syntactic bootstrapping" mode, for the new word "gugle," for example, syntactic knowledge of Agent-Event-Object structure of the sentence "John pushed the gugle" can be used to assign "gugle" to the object of push, rather than "blindly" associating it with all of the possible referents as was done before the SentenceToScene knowledge was acquired.

$$\begin{aligned}&\text{WordToReferent}(i,j)\\&\quad = \text{WordToReferent}(i,j) + \text{OCA}(k,i)\\&\qquad * \text{SEA}(m,j) * \text{Max}(\alpha, \text{SentenceToScene}(m,k)).\end{aligned} \tag{1}$$

3.2. Mapping sentence to meaning

In terms of the architecture in Fig. 3, this mapping can be characterized in the following successive steps. First, words in the Open Class Array are decoded into their corresponding scene referents (via the WordToReferent mapping) to yield

the Predicted Referents Array that contains the translated words while preserving their original order from the OCA

$$\mathrm{PRA}(k,j) = \sum_{i=1}^{n} \mathrm{OCA}(k,i) * \mathrm{WordToReferent}(i,j). \quad (2)$$

Next, each sentence type will correspond to a specific *form to meaning* mapping between the PRA and the SEA, encoded in the SentenceToScene array. The problem will be to retrieve for each sentence type or grammatical form, the appropriate corresponding SentenceToScene mapping.

4. Requirement 3: discriminating between grammatical forms

In the present approach, the first step in discriminating between grammatical structures is to discriminate between open class (e.g., nouns and verbs) and closed class (e.g., determiners and prepositions) words. Newborn infants are sensitive to the perceptual properties that distinguish these two categories (Shi, Werker, & Morgan, 1999), and in adults these categories are processed by dissociable neural systems (Brown, Hagoort, & ter Keurs, 1999). Similarly, artificial neural networks can also learn to make this function/content distinction (Blanc, Dodane, & Dominey, 2003; Morgan, Shi, & Allopenna, 1996). Thus, for the speech input that is provided to the learning model, open and closed class words are directed to separate processing streams that preserve their order and identity, as indicated in Fig. 3.

Given this capability to discriminate between open and closed class words, we are still faced with the problem of using this information to discriminate between different sentence types. To solve this problem, we recall that each sentence type will have a unique constellation of closed class words and/or bound morphemes (Bates et al., 1982) that can be coded in a ConstructionIndex (Eq. (3)) that forms a unique identifier for each sentence type. The ConstructionIndex is a 25 element vector. Each function word is encoded as a single bit in a 25 element FunctionWord vector. When a function word is encountered during sentence processing, the current contents of ConstructionIndex are shifted (with wrap-around) by $n + m$ bits, where n corresponds to the bit that is on in the FunctionWord and m corresponds to the number of open class words that have been encountered since the previous function word (or the beginning of the sentence). Finally, a vector addition is performed on this result and the FunctionWord vector. Thus, the appropriate SentenceToScene mapping for each sentence type can be indexed in ConstructionInventory by its corresponding ConstructionIndex. We have previously demonstrated how a recurrent network can perform this ConstructionIndex function as a form of discrimination between seqeunces of closed class elements (Dominey et al., 2003)

$$\mathrm{ConstructionIndex} = f_{\mathrm{circularShift}}(\mathrm{ConstructionIndex}, \mathrm{FunctionWord}). \quad (3)$$

The link between the ConstructionIndex and the corresponding SentenceToScene mapping is established as follows. As each new sentence is processed, we first reconstruct the specific SentenceToScene mapping for that sentence (Eq. (4)), by mapping words to referents (in PRA) and referents to scene elements (in SEA). The resulting, SentenceToSceneCurrent encodes the correspondence between word order (that is preserved in the PRA Eq. (2)) and thematic roles in the SEA. Note that the quality of SentenceToSceneCurrent will depend on the quality of acquired word meanings in WordToReferent. Thus, syntactic learning requires a minimum baseline of semantic knowledge. Given the SentenceToSceneCurrent mapping for the current sentence, we can now associate this mapping in the ConstructionInventory with the corresponding function word configuration or ConstructionIndex for that sentence, expressed in Eq. (5). In Eqs. (5) and (6), SentenceToScene is linearized for simplification of the matrix multiplication

$$\mathrm{SentenceToSceneCurrent}(m,k) = \sum_{i=1}^{n} \mathrm{PRA}(k,i) * \mathrm{SEA}(m,i), \quad (4)$$

$$\begin{aligned}\text{ConstructionInventory}(i,j) &= \text{ConstructionInventory}(i,j) \\ &\quad + \text{ConstructionIndex}(i) \\ &\quad * \text{SentenceToSceneCurrent}(j). \end{aligned} \tag{5}$$

Finally, once this learning has occurred, for new sentences we can now extract the SentenceToScene mapping from the learned ConstructionInventory by using the ConstructionIndex as an index into this associative memory, illustrated in Eq. (6).

To accommodate the dual scenes for complex events, Eqs. (4)–(7) are instantiated twice each, to represent the two components of the dual scene. In the case of simple scenes, the second component of the dual scene representation is null. This extension is illustrated with an example in Fig. 4

$$\begin{aligned}\text{SentenceToScene}(i) = \sum_{i=1}^{n} &\text{ConstructionInventory}(i,j) \\ &* \text{ConstructionIndex}(j). \end{aligned} \tag{6}$$

We evaluate performance of the model by using the WordToReferent and SentenceToScene knowledge to construct for a given input sentence the "predicted scene". That is, the model will construct an internal representation of the scene that should correspond to the input sentence. This is achieved by first converting the Open-Class-Array into its corresponding scene items in the Predicted-Referents-Array as specified in Eq. (2). The referents are then re-ordered into the proper scene representation via application of the SentenceToScene transformation as described as

$$\text{PSA}(m,i) = \text{PRA}(k,i) * \text{SentenceToScene}(m,k). \tag{7}$$

When learning has proceeded correctly, the PSA contents should match those of the SEA that is directly derived from input to the model. We then quantify performance error in terms of the number of mismatches between PSA and SEA.

5. Experimental results

Hirsh-Pasek and Golinkoff (1996) indicate that children use knowledge of word meaning to acquire a fixed Subject Verb Object template around 18 months, then expand this to non-canonical sentence forms (i.e., those whose word order deviates from the "canonical" SVO, such as the passive which is OVS) at around 24+ months. Tomasello (1999) indicates that fixed grammatical constructions will be used initially and that these will then provide the basis for the development of more generalized constructions (Goldberg, 1995). The following experiments attempt to reproduce particular steps in this type of developmental progression. Training yields in changes in the associative WordToReferent mappings encoding the lexicon, and changes in the ConstructionInventory encoding the form to meaning mappings, indexed by the ConstructionIndex. The ability to handle non-canonical forms with the same architecture that is initially used with the canonical forms in acquisition of the lexicon is an example of how the model simulates and indeed relies on developmental progression.

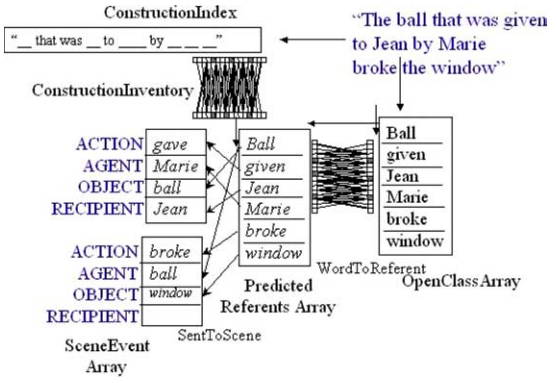

Fig. 4. Representation of complex (relativised) ⟨sentence, meaning⟩ mappings. Relativised sentences contain two complete events (see construction types 5–10 in Section 5.3). By adding a second SceneEventArray, and a second SentenceToScene mapping, the model can accommodate relativised sentences.

5.1. Learning of active forms for simple events

Here, we illustrate a variety of different grammatical construction types, with example sentences for each.

1. Active: The block pushed the triangle.

⟨"*agent verb object*", verb(agent, object)⟩

2. Dative: The block gave the triangle to the moon.
⟨"*agent verb object* to *recipient*", verb(agent, object)⟩

For this first experiment, 17 ⟨sentence, meaning⟩ pairs were generated that employed the five different events, and narrations in the active voice, corresponding to the grammatical forms illustrated in sentences 1 and 2. Example sentences 1–10 are accompanied by their corresponding grammatical constructions in which the nouns and verbs can be substituted with new values to create new sentences of the same type. The model was trained for 32 epochs with the 17 ⟨sentence, meaning⟩ pairs for a total of 544 ⟨sentence, meaning⟩ pairs. During the first 200 ⟨sentence, meaning⟩ pair trials, α in Eq. (1) was 1 (i.e., no syntactic bootstrapping before syntax is acquired), and thereafter it was 0. This was necessary in order to avoid the effect of syntactic random knowledge on semantic learning in the initial learning stages. Performance was measured by comparing the "meaning" that the model generated from an input sentence with the actual meaning provided in the ⟨sentence, meaning⟩ pair. Generalization was tested using new, untrained ⟨sentence, meaning⟩ pairs. The trained system displayed error free performance for all 17 sentences and generalization to new sentences that had not previously been tested. These error free results are obtained under ideal noise free conditions. We have observed that the introduction of processing noise (that more closely resembles the child's environment) still allows effective learning, with graceful degradation proportional to the noise (Dominey & Inui, 2004).

5.2. Passive forms

This experiment examined learning active and passive grammatical forms, employing grammatical forms 1–4. Word meanings were used from Experiment 5.1, so only the structural SentenceToScene mappings were learned.

3. Passive: The triangle was pushed by the block.
⟨"*object* was *verb* by *agent*", verb(agent, object)⟩
4. Dative Passive: The moon was given to the triangle by the block.
⟨"*object* was *verb* to *recipient* by *agent*", verb(agent, object)⟩

17 new ⟨sentence, meaning⟩ pairs were generated with active and passive grammatical forms for the narration. Within three training epochs with the 17 sentences (51 ⟨sentence, meaning⟩ pairs), error free performance was achieved, with confirmation of error free generalization to new untrained sentences of these types. The rapid learning indicates the importance of lexicon in establishing the form to meaning mapping for the grammatical constructions.

5.3. Relative forms for complex events

Here, we consider complex scenes narrated by relative clause sentences and their processing by the model as illustrated in Fig. 4. 11 complex ⟨sentence, meaning⟩ pairs were generated with narration corresponding to the grammatical forms indicated in 5–10:

5. The block that pushed the triangle touched the moon.
⟨"*agent* that *verb1 object* verb2 *object2*", verb1(agent, object1), verb2(agent, object2)⟩
6. The block pushed the triangle that touched the moon.
⟨"*agent2 verb2 agent1* that verb1 *object1*", verb1(agent1, object1), verb2(agent2, agent1)⟩
7. The block that pushed the triangle was touched by the moon.
⟨"*agent1* that *verb1 object1* was verb2 by *agent2*", verb1(agent1, object1), verb2(agent2, agent1)⟩
8. The block pushed the triangle that was touched by the moon.
⟨"*agent2 verb2 object1* that was verb1 by *agent1*", verb1(agent1, object1), verb2(agent2, object1)⟩
9. The block that was pushed by the triangle touched the moon.
⟨"*object1* that was *verb1* by *agent1* verb2 *object2*", verb1(agent1, object1), verb2(object1, object2)⟩

10.
The block was pushed by the triangle that touched the moon.

⟨"*object2* was *verb2* by *agent1* that verb1 *object1*", verb1(agent1, object1), verb2(agent1, object2)⟩

After presentation of 88 ⟨sentence, meaning⟩ pairs, the model performed without error for these six grammatical forms, and displayed error-free generalization to new sentences that had not been used during the training for all six grammatical forms.

5.4. Combined test with and without lexicon

The objective of the final experiment was to verify that the model was capable of learning the 10 grammatical forms together in a single learning session. A total of 27 ⟨sentence, meaning⟩ pairs, used in Experiments 5.2 and 5.3, were employed that exercised the ensemble of 10 grammatical forms. After exposure to six presentations of the 27 ⟨sentence, meaning⟩ trials, the model performed without error. Likewise, in the generalization test the learned values were fixed, and the model demonstrated error-free performance on new sentences, for all 10 grammatical forms, that had not been used during the training.

The rapid acquisition of the grammatical constructions in the presence of pre-learned WordToReferent knowledge is quite striking, and indicates the power of semantic bootstrapping that uses knowledge of word meaning to understand grammatical structure. To further examine this effect, we re-ran these Experiments 5.1–5.4 without using the WordToReferent knowledge (i.e., word meanings) that had been acquired in Experiment 5.1. In this case, the results were equally striking. The active and passive forms in Experiment 5.2 required more than 90 training epochs to achieve error free performance, vs. 3 when word meanings are provided, and 32 training epochs when only the active forms were employed in Experiment 5.1. Training with the relativised constructions in Experiment 5.3 without pre-learned WordToReferent knowledge failed to converge, as did the combined test in Experiment 5.4. This indicates the importance of acquiring an initial lexicon in the context of simple grammatical constructions, or even single word utterances in order to provide the basis for acquisition of more complex grammatical constructions. This is consistent with the developmental observation that infants initially acquire a restricted set of concrete nouns from which they can bootstrap grammar, and further vocabulary (reviewed in Dominey (2000)).

5.5. Generalization to extended construction set

As illustrated above the model can accommodate 10 distinct form-meaning mappings or grammatical constructions, including constructions involving "dual" events in the meaning representation that correspond to relative clauses. Still, this is a relatively limited size for the construction inventory. We have subsequently demonstrated that the model can accommodate 38 different grammatical constructions that combine verbs with two or three arguments, active and passive forms and relativisation, along with additional sentence types including: conjoined (John took the key and opened the door), reflexive (The boy said that the dog was chased by the cat), and reflexive pronoun (The block said that it pushed the cylinder) sentence types. The consideration of these sentence types requires us to address how their meanings are represented. Indeed, our current scene analysis capabilities do not include the detection of reflexive verbs such as "said" and so the meanings were hand coded for these sentences. Conjoined sentences are represented by the two corresponding events, e.g. *took*(*John*, *key*), *open*(*John*, *door*) for the conjoined example above. Reflexives are represented, for example, as *said*(*boy*), *chased*(*cat*, *dog*). This assumes indeed, for reflexive verbs (e.g. said, saw), that the meaning representation includes the second event as an argument to the first. Finally, for the reflexive pronoun types, in the meaning representation the pronoun's referent is explicit, as in *said*(*block*), *push*(*block*, *cylinder*) for "The block said that it pushed the cylinder."

For this testing, the ConstructionInventory is implemented as a lookup table in which the ConstructionIndex is paired with the corresponding SentenceToScene mapping during a single learning

trial. Based on the tenets of the construction grammar framework (Goldberg, 1995), if a sentence is encountered that has a form (i.e., ConstructionIndex) that does not have a corresponding entry in the ConstructionInventory, then a new construction is defined. Thus, one exposure to a sentence of a new construction type allows the model to generalize to any new sentence of that type. In this sense, developing the capacity to handle a simple initial set of constructions leads to a highly extensible system. Using the training procedures as described above, with a pre-learned lexicon (WordToReferent), the model successfully learned all of the constructions, and demonstrated generalization to new sentences that it was not trained on.

That the model can accommodate these 38 different grammatical constructions with no modifications indicates its capability to generalize to new constructions. That is, the method of forming the ConstructionIndex based on the configuration of closed class words is a reliable method for discriminating between different grammatical constructions, whether the ConstructionIndex is then used in an associative memory or the functionally equivalent lookup table. This translates to a (partial) validation of the hypothesis that across languages, thematic role assignment is encoded by a limited set of parameters including word order and grammatical marking, and that distinct grammatical constructions will have distinct and identifying ensembles of these parameters.

5.6. Extension of the construction framework to spatial relations and attention

The concept of ''emergence'' entails that existing processes can provide the basis for the emergence of new behavioral functionality. We have seen how the construction framework provides a basis for encoding the structural mappings between sentences and meaning in an organized and generalized manner. In theory, this construction framework should extend to analogous cognitive domains. Here, we will investigate how this framework can be extended to the domain of spatial relations. The extension involves two components. First, we should demonstrate that the concept of using simple perceptual primitives which can easily be extracted from the topographic retinal image can be applied to spatial relations as well as physical events for extracting meaning from vision. Second, we should demonstrate that this meaning can be encoded in a predicate-argument format that is compatible with the ''standard'' meaning representation in the grammatical construction learning model.

Quinn et al. (2002) and Quinn (2003) have demonstrated that by the age of 6–7 months, infants can learn binary spatial relations such as left, right, above, below in a generalized manner, as revealed by their ability to discriminate in familiarization-test experiments. That is, they can apply this relational knowledge to scenes with new objects in these spatial relations. In theory, the predicate-argument representation for event structure that we have described above can provide the basis for representing spatial relations in the form Left (X,Y), Above(X,Y), etc., where X is the object that holds the spatial relation with the referent Y. That is, Left(X,Y) corresponds to ''X is left of Y''. In order to extract spatial relations from vision, we return to the visual processing system described above. Based on the observations of Quinn (2003) we can consider that by 6–7 months, the perceptual primitives of Relation(X,Y) are available, where Relation corresponds to Left, Right, Above, and Below. The mapping of sentence structure onto the predicate argument then can proceed as described above for event meaning.

One interesting problem presents itself however, related to referential ambiguity. Fig. 5 illustrates the spatial configuration after a human user has placed the cylinder in its current position and said ''The cylinder is below the triangle''. Given this image, any one of the four objects could be the subject of the relation, and any one of the remaining three could be the referent, thus yielding 12 possible relations. The problem then is one of referential uncertainty, or ''what is the speaker talking about?''

Tomasello (2003) clearly emphasizes the crucial role of shared attention between the speaker and listener in solving this referential uncertainty. One of the most primitive forms of attention is related to the detection of movement, and the act of ''showing'' something almost always involves

either pointing to or moving the object. In this context, Kellman, Gleitman, and Spelke (1987) demonstrated that as early as 16 weeks, infants are sensitive to object motion that can provide the basis for object identification, discriminating retinal displacement due to their own movement from that due to displacement of objects. Thus, we employed a simple attention mechanism based on motion to select the last object in motion (cylinder in the example of Fig. 5) as the target object. Still, the intended referent for the "below" relation could be any one of the multiple other objets, and so the problem of referential ambiguity must still be resolved. We hypothesize that this redundancy is resolved in human interaction based on two perceptual parameters. First, spatial proximity, or distance from the target will be used. That is, the observer will give more attentional preference to relations involving the target object and other objects that are closest to it. The second parameter is the angular "relevance" of the relations, quantified in terms of the angular distance from the cardinal positions *above*, *below*, *left*, and *right*. Fig. 5(b) represents the application of this perceptual attention mechanism that selects the relation Below(Cylinder, Triangle) as the most relevant, revealed by the height of the peak for the triangle in 5B.

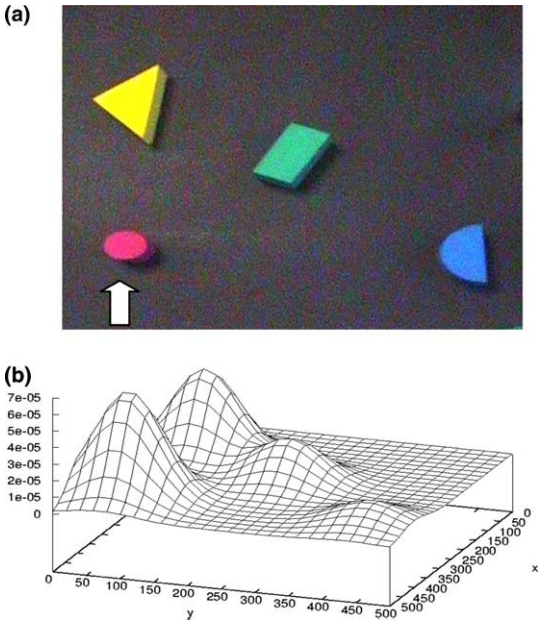

Fig. 5. Spatial attention for relation selection. The human user shows the robot a spatial relation and describes it. How does the robot know which of the multiple relations is the relevant one? (a) The cylinder (lower left) has been moved into its current position, and now holds spatial relations with the three other objects. (b) Based on parameters of (1) minimal distance from the target object and (2) minimal angular distance from the four principal directions (above, below, left, right).. In this case, the most relevant relation (indicated by the height of the two highest peaks) is Below(Cylinder, Triangle).

In order to validate this attentional strategy for extracting spatial relations, we collected data from four human subjects who were instructed to "teach" the robot the spatial relations by demonstrating and narrating spatial relations with the four colored blocks. The resulting data were 74 training examples, each consisting of the short video sequence in which the subject "showed" or demonstrated a spatial relation, and provided the corresponding description of the demonstrated relation. The spatial attention mechanism determined the most relevant spatial relation for the video sequence in each case in order to extract the "meaning" in terms of a spatial relation. Of the resulting 74 meanings that were automatically extracted using this mechanism, 67 (95%) corresponded exactly to the meaning described by the subject, i.e., to the subject's intended meaning.

Fig. 6 illustrates the robustness of the two underlying assumptions with respect to human performance. In Fig. 6(a), we see that the human subjects reliably demonstrated relations in a pertinent manner, adhering closely to the four principal axes. Likewise, Fig. 6(b) illustrates that in the large majority of the examples, subjects placed the target object closer to the referent object than to the other objects in the scene. This demonstrates that perceptual primitives of motion, distance, and angle can be reliably used in order to construct a higher level attention capability.

The 74 resulting ⟨sentence, relation-meaning⟩ pairs were then used as input to the grammatical construction learning model. Fig. 7 illustrates the learning curve for this data. After 15 exposures to data set, the model converges to a stable performance. Of the 74 input ⟨sentence, meaning⟩ pairs, 67 are well formed, and 7 are not well formed, i.e., the extracted relation does not correspond to

the described meaning. After training, the model correctly identifies the 7 non-well-formed ⟨sentence, meaning⟩ pairs, and performs at 91% correct (61/67) for the remaining correct pairs. Interestingly, the misunderstood correct sentences directly followed the erroneous pairs in the training data, indicating that learning with wrong examples has a short-termed impairment effect on subsequent performance. More importantly, we verified that based on training with correct examples, the model could generalize this knowledge to a new ⟨sentence, relation-meaning⟩ generalization data set.

5.7. Semantic compositionality and the emergence of the "between" relation

The previous section demonstrated that perceptual primitives could be used to extract relevant spatial relations from visual scenes and that the grammatical construction learning model could accommodate these spatial relations for learning ⟨sentence, relation⟩ pairs. While this demonstrates that the initial system could extend to learning new types of meaning (i.e., spatial relations), it does not really get at the emergence of new behavior.

Interestingly, this issue can be approached from the perspective of Quinn's observation of the development of the relation "between". The ability to discriminate "between" occurs significantly later than that for above–below and left–right (9 months vs 6 months, respectively) (Quinn, Adams, Kennedy, Shettler, & Wasnik, 2003). This suggests that the ternary relation "between" is more complex than the binary relations above–below and left–right. It also suggests that between could in fact be constructed from these more primitive relations.

This provides a potentially interesting scenario for emergence. As illustrated in Fig. 4, the model is capable of representing two predicate-argument meanings in parallel. This was initially provided to account for dual event meanings in relativised sentences, but can in fact be used for arbitrary predicate-argument structures. In a spatial array

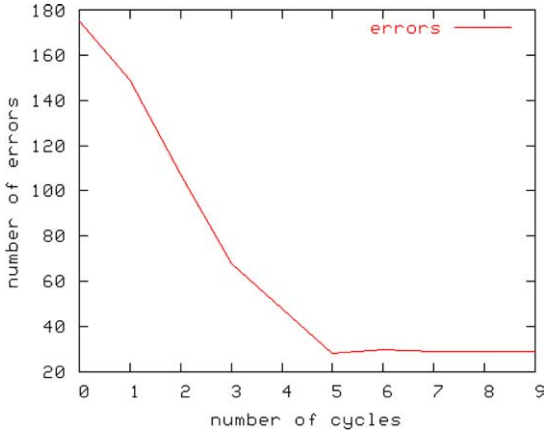

Fig. 7. Learning performance of grammatical construction learning model with the 74 relational training examples. Each cycle (epoch) corresponds to a full pass through the 74 ⟨sentence, relation⟩ pairs. Final errors are due to incorrect ⟨sentence, meaning⟩ data in the input.

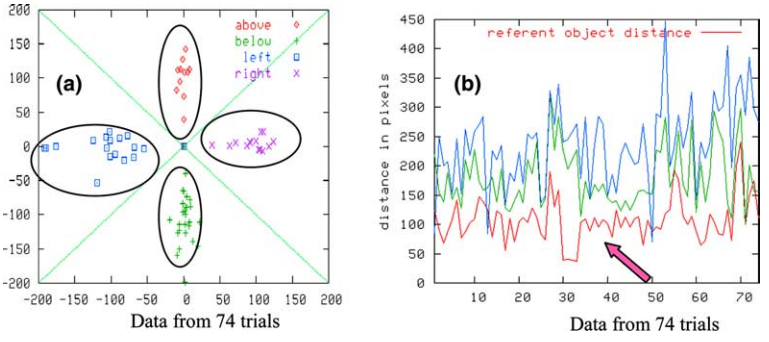

Fig. 6. (a) Location of the target with respect to referent object in the Relation(target, referent) relations. Note that the experimental subjects place the target object closely aligned with appropriate direction (left, right, above, below), and not ambiguously. (b) Distance between target and other objects. Lowest curve is for the intended referent, extracted from the verbal descriptions. As predicted, subjects almost invariably place the target closest to the intended referent.

X–Y–Z, in which Y is between X and Z, the composite relation "between" can be decomposed into the relations Y is left of Z and Y is right of X. As illustrated in Fig. 8, these two relations can be encoded in the two SceneEventArrays, and the system should then map the sentence "Y is between X and Z" onto this dual meaning structure.

In order to test this capability, we "showed" the between relation to the robot in 10 trials. In each case, the "between" object was placed between the two referent objects, and the corresponding description was provided. We modified the spatial attention mechanism to provide the two most relevant relations (defined as described above as a function of proximity and angular pertinence), rather than the single most relevant. In the 10 training trials, the two resulting relations were always of the form Left(Y,Z) and Right(Y,X). The model was then trained with the 10 corresponding ⟨X is between Y and Z; (Left(Y,Z), Right(Y,X))⟩ complex pairs, and two learning results were observed. First, from the lexical level, the word "between" became ambiguously associated with the "meanings" left and right. Second, from the phrasal construction level, the construction "X is between Y and Z" became correctly associated with the mapping to Left(Y,Z) and Right(Y,X), as illustrated in Fig. 8(a).

While these initial results are encouraging, they also raise important issues of limitations. First, the mapping of lexical item "between" onto meanings "left" and "right" is not ideal. Second, if the relations are extracted in the order Right(Y,X) and Left(Y,Z) (rather than the reverse) then the system will not recognize this as between, although it is spatially equivalent. The solution to both of these problems is the introduction of an intermediate level of representation in which primitives can be combined to form composite relations that are directly mapped into their constructions. This type of architectural change would provide the basis for an open-ended flexibility for the learning of diverse composite relational structures including spatial structures (e.g., arch, stack, "T", etc.) as well as temporal and logical structures (if–then, before, after, because, etc.). These learning capabilities will be explored in our future research.

6. Discussion

Already at birth, infants are sensitive to the prosodic structure of language that allows them to perform the first crucial discrimination between content and function words in acquiring the structure of their language (Shi et al., 1999). Indeed, we have demonstrated that a temporal recurrent network of leaky integrator neurons is sensitive to the temporal structure of language (Dominey & Ramus, 2000) and can perform lexical categorization of open and closed class words (Blanc et al., 2003). At the same time during the first year of life, the infant begins to construct meaning from the perceptual world (Mandler, 1999) exploiting per-

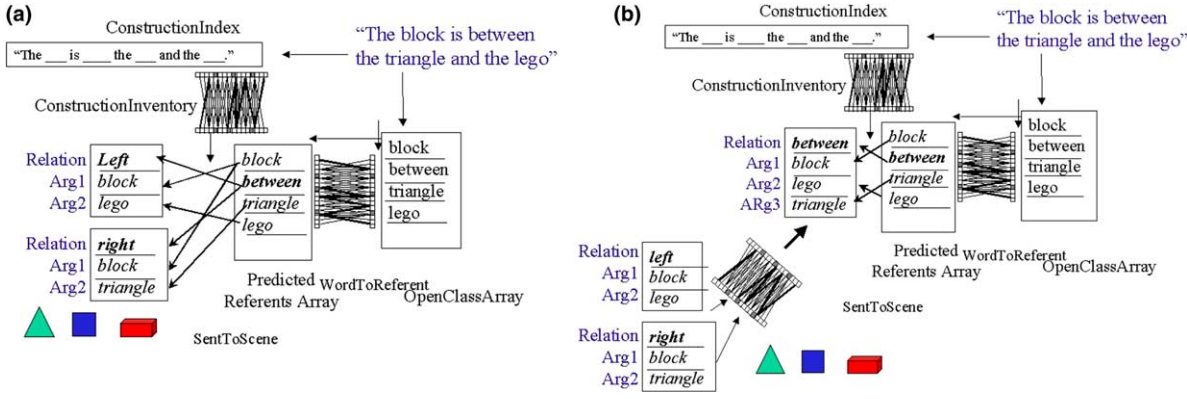

Fig. 8. (a) Representation of "BETWEEN" in the existing architecture. (b) More general representation for allowing the construction of arbitrary composite relations.

ceptual primitives including force dynamic properties such as contact, support, and attachment (Talmy, 1988) in order to construct meaning in terms of physical events (Kotovsky & Baillargeon, 1998; Mandler, 1999).

In this context, the current work illustrates how computer vision systems are now able to exploit such physical regularities in order to form predicate-argument descriptions of visual scenes (see also Siskind, 2001; Steels & Baillie, 2002). While this level of analysis concentrates on physical events, it should extend to more open ended semantics with respect to time, causation and agency, to the extent that these aspects are composed from or encoded in the physical event structure.

Combined with learning mechanisms that exploit cross-situational statistics of word meanings (Siskind, 1996) and the mapping between grammatical structure and event structure, as illustrated here, learning systems can move from word to sentence in language acquisition. There, the synergy between word learning that allows syntactic structure to be revealed, and the syntactic structure which in turn facilitates new word acquisition allows for a rapid learning capability. The current research links these elements together in a grounded robotic platform for the study of language acquisition and comprehension. Interestingly, these results are obtained with largely domain general mechanisms for learning associative mappings and structural transformations. Specifically, what is innate is: (1) an associative memory (WordToReferent) that associates words to referents, (2) a structure mapping mechanism (SentenceToScene) that performs a structural transformation from sentence to meaning, (3) an associative memory (ConstructionInventory) that allows the storage and retrieval of these transformations, (4) a form of recurrent network for processing closed class elements (grammatical function words) that acts as an index (ConstructionIndex) into this memory, and finally, (5) a discrimination mechanism that separates elements into the open or closed class processing stream. All of these functions have been implemented as biologically plausible neural networks (Dominey et al., 2003), and none of these functions is inherently language-related. What makes them become language related is their organization in the specified configuration, including the perceptual grounding. This is in striking contrast to propositions of a Universal Grammar that is by definition modular and fully language specific.

With respect to this language processing, the system demonstrates an interesting, though limited generalization capability. Once a given grammatical construction has been learned (e.g., one of the 10 numbered constructions in Section 5), that construction can then be used to generalize in a systematic manner to all new sentences built from that construction. The ''slots'' in the construction are simply filled in with new nouns and verbs. This explains the error free generalization to new sentences. The limitation is that the system must first be exposed to a well-formed construction (⟨sentence, meaning⟩ pair) in order to learn the defining structural mapping.

Similarly, the extraction of events from the dynamic scenes was ''hard coded'' in a parser that looked for specific categories of contact sequences that correspond to push, touch, take, and give (Fig. 2). A more realistic approach would be to extract only the true primitive – contact – and then learn to associate structured sequences of contacts with the corresponding event types. This would allow much more flexibility in the extraction of meaning. Interestingly, this issue is partially addressed in Section 5.7 (see also Fig. 8(b)), in which a structured ensemble of perceptual–semantic primitives becomes associated with a composite semantic structure. Again, this direction will be pursued in our future research.

A related limitation concerns the syntactic complexity and compositionality. Miikkulainen (1996) demonstrated an elegant and efficient mechanism for handling relativised sentences in a hybrid neural network. In order to do so, the system required two additional capabilities that are not in the current model, including a hard coded parser, and a stack. The point of interest in the current approach is that ''hierarchical'' sentences are processed without a stack, relying on the semantic representation of the meaning as the only data structure required. Likewise, Miikkulainen's hard coded parser is replaced by a system that learns sentence form to

meaning mappings. The advantage is in the extreme simplicity of the system, and the price is the limit in compositionality. That is, rather than using a fully generalized recursive stack-based parsing strategy, our system uses a form of "shallow" parsing that directly maps sentences with recursive structure onto their corresponding meaning representations without explicitly performing recursion. Interestingly, recent human experimental studies suggest that at least in certain conditions, humans tend to rely on these simpler mechanisms (Sanford & Sturt, 2002). Our future work in this context will examine how embedded noun phrases in relativised sentences can be extracted via statistical learning methods, with the goal of producing a learned compositionality capability when existing constructions are not sufficient.

We note that the experiments presented here employed a form of batch learning in which multiple exposures to ⟨sentence, meaning⟩ pairs were presented to the system. In a sense, this is unsatisfactory in that it does not reflect on-line human–robot interaction. Indeed, the ⟨sentence, meaning⟩ pairs could have been generated and presented in real-time in a functionally equivalent manner, though this would have produced excessively long interaction sessions with respect to the humans. However, once the lexical and construction knowledge has been acquired, we have now demonstrated how it can be used for quite satisfactory real-time human–robot interaction (Dominey, Boucher, & Inui, 2004).

In conclusion, the current study demonstrates: (1) that the perceptual primitive of contact (available to infants at 5 months) can be used to perform event description in a manner that is similar to but significantly simpler than Siskind (2001), and can be extended to accommodate spatial relation encoding, (2) that a novel implementation of principles from construction grammar can be used to map sentence form to these meanings together in an integrated system, (3) that relative clauses can be processed in a manner that is similar to, but requires less specific machinery (e.g. no stack) than that in Miikkulainen (1996), and finally (4) that the resulting system displays robust acquisition behavior that reproduces certain observations from developmental studies with very modest "innate" language specificity.

Acknowledgements

Supported by the OHLL, EuroCores OMLL, French ACI Integrative and Computational Neuroscience, and HFSP MCILA Projects.

References

Bates, E., McNew, S., MacWhinney, B., Devescovi, A., & Smith, S. (1982). Functional constraints on sentence processing: a cross linguistic study. *Cognition*(11), 245–299.

Blanc, J. M., Dodane, C., & Dominey, P. F. (2003). Temporal processing for syntax acquisition: a simulation study. *25th annual meeting of the cognitive science society*. Boston, MA, USA.

Brown, C. M., Hagoort, P., & ter Keurs, M. (1999). Electrophysiological signatures of visual lexical processing: open- and closed-class words. *Journal of Cognitive Neuroscience, 11*(3), 261–281.

Chomsky, N. (1995). *The minimalist program*. MIT.

Chang, N. C., & Maia, T. V. (2001). Grounded learning of grammatical constructions. In *AAAI spring symposium on learning grounded representations, Stanford, CA*.

Cottrel, G. W., Bartell, B., & Haupt, C. (1990). Grounding meaning in perception. In *Proceedings of the GWAI90, 14th German workshop on artificial intelligence* (pp. 307–321). Berlin, New York: Springer.

Dominey, P. F., & Ramus, F. (2000). Neural network processing of natural lanugage: I. Sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes, 15*(1), 87–127.

Dominey, P. F. (2000). Conceptual grounding in simulation studies of language acquisition. *Evolution of Communication, 4*(1), 57–85.

Dominey, P. F., Boucher, J. D., & Inui, T. (2004). Building an adaptive spoken language interface for perceptually grounded human–robot interaction. In *Proceedings of the IEEE-RAS/RSJ international conference on humanoid robots*.

Dominey, P. F., Hoen, M., Lelekov, T., & Blanc, J. M. (2003). Neurological basis of language in sequential cognition: evidence from simulation, aphasia and ERP studies. *Brain and Language, 86*(2), 207–225.

Dominey, P. F., & Hoen, M. (2005). Structure mapping and semantic integration in a construction-based neurolinguistic model of sentence processing, *Cortex*, In press.

Dominey, P. F., & Inui, T. (2004). A developmental model of syntax acquisition in the construction grammar framework with cross-linguistic validation in English and Japanese. In

Proceedings of the 20th international conference on computational linguistics, psycho-computational models of human language acquisition, Geneva. pp. 33–40.

Elman, J. (1990). Finding structure in time. *Cognitive Science, 14*, 179–211.

Feldman, J. A., Lakoff, G., Stolcke, A., & Weber, S. H. (1990). Miniature language acquisition: a touchstone for cognitive science. In *Proceedings of the 12th Annual Conference on Cognitive Science Society* (pp. 686–693). Cambridge, MA: MIT.

Feldman, J., Lakoff, G., Bailey, D., Narayanan, S., Regier, T., & Stolcke, A. (1996). L0: The First Five Years. *Artificial Intelligence Review, 10*, 103–129.

Goldberg, A. (1995). *Constructions*. Chicago and London: Univ. Chicago Press.

Hirsh-Pasek, K., & Golinkoff, R. M. (1996). *The origins of grammar: evidence from early language comprehension*. Boston: MIT Press.

Hoen, M., Pachot-Clouard, M., Segebarth, C., & Dominey, P. F. (2005). When Broca experiences the Janus syndrome. An er-fMRI study comparing sentence comprehension and cognitive sequence processing. *Cortex*, In press.

Kellman, P. J., Gleitman, H., & Spelke, E. S. (1987). Object and observer motion in the perception of objects by infants. *Journal of Experimental Psychology – Human Perception and Performance, 13*(4), 586–593.

Kotovsky, L., & Baillargeon, R. (1998). The development of calibration-based reasoning about collision events in young infants. *Cognition, 67*, 311–351.

Langacker, R. (1991). *Foundations of cognitive grammar. Practical applications* (Vol. 2). Stanford: Stanford University Press.

Mandler, J. (1999). Preverbal representations and language. In P. Bloom, M. A. Peterson, L. Nadel, & M. F. Garrett (Eds.), *Language and space* (pp. 365–384). MIT Press.

Miikkulainen, R. (1996). Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science, 20*, 47–73.

Morgan, J. L., Shi, R., & Allopenna, P. (1996). Perceptual bases of rudimentary grammatical categories. In J. L. Morgan & K. Demuth (Eds.), *Signal to syntax* (pp. 263–286). Mahwah, NJ: Lawrence Erlbaum.

Quinn, P. C. (2003). Concepts are not just for objects: categorization of spatial relation information by infants. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development: making sense of the blooming, buzzing confusion* (pp. 50–76). Oxford University Press.

Quinn, P. C., Adams, A., Kennedy, E., Shettler, L., & Wasnik, A. (2003). Development of an abstract category representation for the spatial relation between in 6- to 10-month-old infants. *Developmental Psychology*(39), 151–163.

Quinn, P. C., Polly, J. L., Furer, M. J., Dobson, V., & Nanter, D. B. (2002). Young infants' performance in the object-variation version of the above–below categorization task. *Infancy, 3*, 323–347.

Roy, D., & Pentland, A. (2002). Learning words from sights and sounds: a computational model. *Cognitive Science, 26*(1), 113–146.

Sanford, A. J., & Sturt, P. (2002). Depth of processing in language comprehension: not noticing the evidence. *Trends in Cognitive Sciences, 6*(9), 382–386.

Shi, R., Werker, J. F., & Morgan, J. L. (1999). Newborn infants' sensitivity to perceptual cues to lexical and grammatical words. *Cognition, 72*(2), B11–B21.

Siskind, J. M. (1996). A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*(61), 39–91.

Siskind, J. M. (2001). Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of AI Research*(15), 31–90.

Steels, L. (2001). Language games for autonomous robots. In *IEEE Intelligent Systems* (Vol. 16, No. 5, pp. 16–22). New York: IEEE Press.

Steels, L., & Baillie, JC. (2002). Shared grounding of event descriptions by autonomous robots. *Robotics and Autonomous Systems, 43*(2–3), 163–173.

Stolcke, A., & Omohundro, S. M. (1994). Inducing probablistic grammars by Bayseian model merging. In *Proceedings of the 2nd international colloquial on grammatical inference. Grammatical inference and applications*. Springer.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science, 10*(2), 117–149.

Tomasello, M. (1999). The item-based nature of children's early syntactic development. *Trends in Cognitive Science, 4*(4), 156–163.

Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Cambridge: Harvard University Press.