In Berthouze, L., Kozima, H., Prince, C. G., Sandini, G., Stojanov, G., Metta, G., and Balkenius, C. (Eds.)
Proceedings of the Fourth International Workshop on Epigenetic Robotics
Lund University Cognitive Studies, 117, ISBN 91-974741-3-4

Developmental Stages of Perception and Language Acquisition in a Perceptually Grounded Robot

Peter Ford Dominey (dominey@ isc.cnrs.fr), Jean-David Boucher (boucher@ isc.cnrs.fr)
Institut des Sciences Cognitives, CNRS
67 Blvd. Pinel, 69675 Bron Cedex, France
http://www.isc.cnrs.fr/dom/dommenu-en.htm

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. Mapping of sentence form to meaning is performed by learning grammatical constructions that are retrieved from a construction inventory based on the constellation of closed class items 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.

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 alternative, 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, attachment (Talmy 1988) the infant could construct progressively more elaborate representations of visuospatial meaning. Thus, the physical event "collision" is a form of the perceptual primitive "contact". Kotovsky & 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 et al. (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, 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 (2003). The use of these intermediate representations renders the system 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). These constructions are initially rather fixed and specific, and later become generalized into a more abstract compositional form employed by the adult (Tomasello 1999). 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 et al. 1990, 1996; Langacker 1991; Mandler 1999; Talmy 1998).

Figure 1. Perceptually grounded robotic system

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; Langacker 1991; Goldberg 1995), Chang & Maia (2001) exploited the relations between action representation and simple verb frames in a construction grammar approach, and Cottrell et al. (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 and 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 et al. 1982). That model was 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 et al. 2003).

1.2 Objectives

The goals of the current study are fourfold: First to test the hypothesis that meaning 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 determine whether the model of Dominey (2000) can be extended to handle embedded relative clauses; Third 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 Context

As illustrated in Figure 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 system (IBM ViaVoice^{TM}). 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. A total of ~300 <sentence, scene> pairs were tested in the following experiments.

2. Requirement 1: Extracting Meaning

For a given video sequence (see snapshot in Figure 2) the visual scene analysis generates the corresponding

event description in the format *event(agent, object, recipient)*.

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 element that had a larger relative velocity towards the other element involved in the contact. Based on these parameters of contact, scene events are recognized as follows:

Touch(agent, object): A single contact, in which (a) the duration of the contact is inferior to *touch_duration* (1.5 seconds), 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 sec), 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(a,o), 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 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).

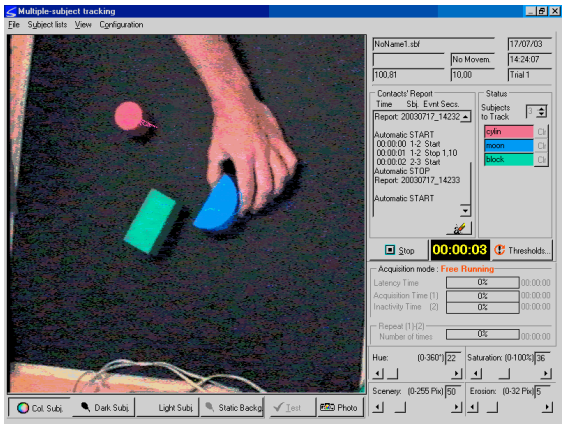

Figure 2. Snapshot of scene event processing.

2.2 Complex "Hierarchical" Events

The events described above are simple in the sense that there 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. 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 grammatical function words or morphemes (Bates et al. 1982).

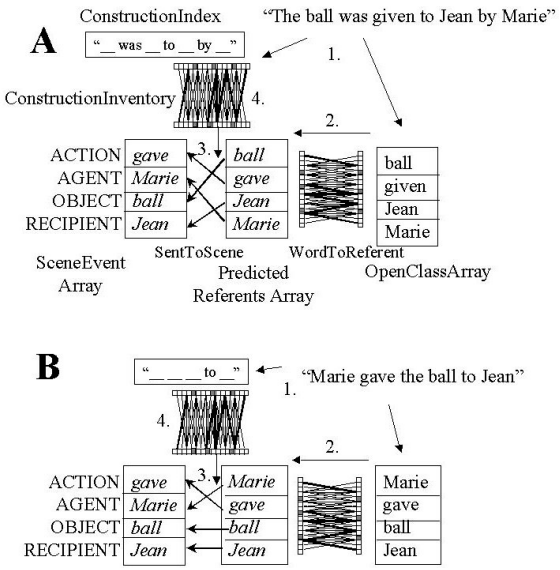

Figure 3. Model Overview: Processing of active and passive sentence types in A, B, respectively. On input, Open class words populate the Open Class Array (OCA), and closed class words populate the Construction index. Visual Scene Analysis populates the Scene Event Array (SEA) with the extracted meaning as scene elements. Words in OCA are translated to Predicted Referents via the WordToReferent mapping to populate the Predicted Referents Array (PRA). PRA elements are mapped onto their roles in the Scene Event Array (SEA) by the SentenceToScene mapping, specific to each sentence type. 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.

The mapping of sentence form onto meaning (Goldberg 1995) 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 & Pentland (2000) 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.

3.1 Word Meaning

In the initial learning phases there is no influence of syntactic knowledge and the word-referent associations are stored in the WordToReferent matrix (Eqn 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 (Eqn. 4, 5 & 6 below). 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 Eqn. 1. In this "syntactic bootstrapping" 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.

$$\text{WordToReferent}(i,j) = \text{WordToReferent}(i,j) + \text{OCA}(k,i) * \text{SEA}(m,j) * \text{Max}(\alpha, \text{SentenceToScene}(m,k)) \quad (1)$$

Indices: k(1:6) - words; m(1:6) – scene elements; i(1:25), j(1:25) – elements in word and scene item vectors, respectively.

3.2 Mapping Sentence to Meaning

In terms of the architecture in Figure 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 (Eqn 2).

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

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, the appropriate corresponding SentenceToScene mapping.

4. Requirement 3: Discriminating Between Grammatical Forms

The first step in discriminating between grammatical structures is to discriminate between open class (e.g. nouns, verbs) and closed class (e.g. determiners, prepositions) words. Newborn infants are sensitive to the perceptual properties that distinguish these two categories (Shi et al. 1999), and in adults these categories are processed by dissociable neural systems (Brown et al. 1999). Similarly, artificial neural networks can also learn to make this function/content distinction (Morgan et al. 1996, Blanc et al. 2003). 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 Figure 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 (Eqn.3) that forms a unique identifier for each sentence type, shifting the current contents by the index of the ON bit in FunctionWord, then ANDing the FunctionWord vector. The appropriate SentenceToScene mapping for each sentence type can be indexed in ConstructionInventory by its corresponding ConstructionIndex.

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

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 (Eqn 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 Eqn 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 it in the ConstructionInventory with the corresponding function word configuration or ConstructionIndex for that sentence, expressed in (Eqn 5). In Eqns 5, 6 SentenceToScene is linearized for simplification.

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

$$\text{ConstructionInventory}(i,j) = \text{ConstructionInventory}(i,j) + \text{ConstructionIndex}(i) * \text{SentenceToSceneCurrent}(j) \quad (5)$$

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 Eqn. 6.

To accommodate the dual scenes for complex events Eqns. 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.

$$\text{SentenceToScene}(i) = \sum_{i=1}^{n} \text{ConstructionInventory}(i,j) * \text{ConstructionIndex}(j) \quad (6)$$

We evaluate performance 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 Eqn. 2. The referents are then re-ordered into the proper scene representation via application of the SentenceToScene transformation as described in Eqn. 7.

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

When learning has proceeded correctly, the predicted scene array (PSA) contents should match those of the scene event array (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 & Golinkoff (1996) indicate that children use knowledge of word meaning to acquire a fixed SVO template around 18 months, then expand this to non-canonical sentence forms 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 follow this type of developmental progression. Training results 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.

5.1 Learning of Active Forms for Simple Events

1. Active: The block pushed the triangle.
2. Dative: The block gave the triangle to the moon.

For this experiment, 17 scene/sentence pairs were generated that employed the 5 different events, and narrations in the active voice, corresponding to the grammatical forms 1 and 2. The model was trained for 32 passes through the 17 scene/sentence pairs for a total of 544 scene/sentence pairs. During the first 200 scene/sentence pair trials, α in Eqn. 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 random effect of syntactic knowledge on semantic learning in the initial learning stages. The trained system displayed error free performance for all 17 sentences, and generalization to new sentences that had not previously been tested.

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.
4. Dative Passive: The moon was given to the triangle by the block.

Seventeen new scene/sentence pairs were generated with active and passive grammatical forms for the narration. Within 3 training passes through the 17 sentences (51 scene/sentence 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. Eleven complex scene/sentence pairs were generated with narration corresponding to the grammatical forms indicated in 5 – 10:

5. The block that pushed the triangle touched the moon.
6. The block pushed the triangle that touched the moon.
7. The block that pushed the triangle was touched by the moon.
8. The block pushed the triangle that was touched the moon.
9. The block that was pushed by the triangle touched the moon.
10. The block was pushed by the triangle that touched the moon.

After presentation of 88 scene/sentence pairs, the model performed without error for these 6 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 scene/sentence pairs, used in Experiments 5.2 and 5.3, were employed that exercised the ensemble of 10 grammatical forms. After exposure to 6 presentations of the 27 scene/sentence 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 ten 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 it 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 5.2 required more than 90 Training Passes to achieve error free performance, vs. 3 Training Passes when word meanings are provided, and 32 Training Passes when only the active forms were employed in 5.1. Training with the relativised constructions in 5.3 without pre-learned WordToReferent knowledge failed to converge, as did the combined test in 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 relativization, 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. 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. 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

Part of the “emergence” framework holds 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. Quinn et al (2002) 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 et al. (2002) 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.

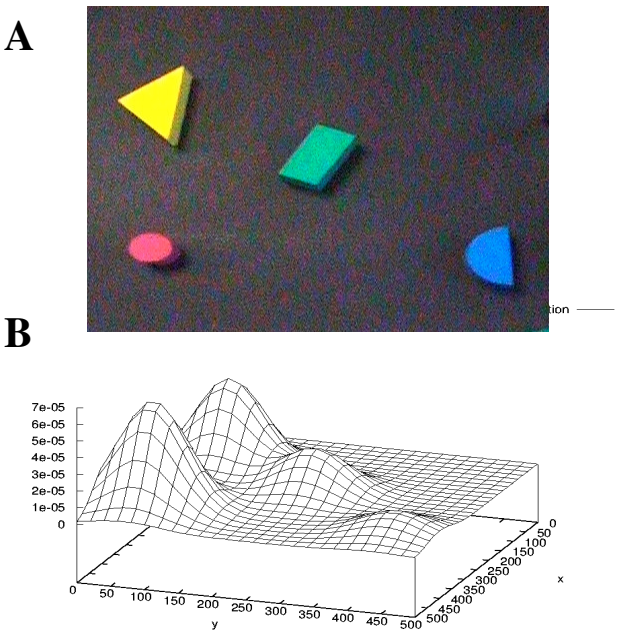

Figure 4. Spatial Attention for Relation Selection. The human user shows the robot a spatial relation and describes it. How does the robot know which of the multiple relations is the relevant one? A. The cylinder (lower left) has been moved into its current position, and now holds spatial relations with 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).

Figure 4 illustrates the spatial configuration after a human user has placed the cylinder in its current position and said "The cylinder is below the triangle". A simple attention mechanism based on motion is used to select the cylinder as the target object, but the intended referent for the "below" relation could be any one of the multiple other objets, and so the problem of referential ambiguity must be resolved. We hypothesize that this redundancy is resolved based on two perceptual parameters. First, spatial proximity 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. Figure 4B 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 4B.

We collected data training data in which a human observer demonstrated and narrated spatial relations with the four objects. The spatial attention mechanism extracted for each case the most relevant spatial relation, and the resulting <sentence, relation-meaning> pairs were used for training in the same procedure as in condition A for active sentences and simple events. The model demonstrated successful learning of the four object names and the four spatial relation terms, and could generalize this knowledge to a new <sentence, relation-meaning> generalization data set.

6. Discussion: Successive Emergence

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 and 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 perceptual primitives including force dynamic properties such as contact, support and attachment (Talmy 1988) in order to construct meaning in terms of physical events (Mandler 1999, Kotovsky & Baillargeon 1998). As illustrated here 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 and Baillie 2003).

Combined with learning mechanisms that exploit cross-situational statistics meanings of words (Siskind 1996), and the mapping between grammatical structure and event structure, as illustrated here, learning systems can move from word to sentence in langage 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 (see Dominey 2000). The current research links these elements together in a grounded robotic platform for the study of language acquisition and comprehension.

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.

Acknowledgments

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 JM, Dodane C, Dominey PF (2003) Temporal processing for syntax acquisition: A Simulation Study, In Press, *Proceedings of the 25th Ann Conf. Cog. Sci. Soc.*, MIT, Cambridge MA

Brown CM, 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 NC, Maia TV (2001) Grounded learning of grammatical constructions, *AAAI Spring Symp. On Learning Grounded Representations*, Stanford CA.

Cottrel GW, Bartell B, Haupt C. (1990) Grounding Meaning in Perception. In Proc. GWAI90, 14th German Workshop on Artificial Intelligence, pages 307--321, Berlin, New York,. Springer Verlag.

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

Dominey PF (2000) Conceptual Grounding in Simulation Studies of Language Acquisition, *Evolution of Communication*, 4(1), 57-85.

Dominey PF, Hoen M, Lelekov T, Blanc JM (2003) Neurological basis of language in sequential cognition: Evidence from simulation, aphasia and ERP studies, *Brain and Language,* 86(2):207-25

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

Feldman JA, Lakoff G, Stolcke A, Weber SH (1990) Miniature language acquisition: A touchstone for cognitive science. In *Proceedings of the 12th Ann Conf. Cog. Sci. Soc.* 686-693, MIT, Cambridge MA

Feldman J., G. Lakoff, D. Bailey, S. Narayanan, T. Regier, A. Stolcke (1996). L0: The First Five Years. *Artificial Intelligence Review*, v10 103-129.

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

Hirsh-Pasek K, Golinkof RM (1996) *The origins of grammar: evidence from early language comprehension.* MIT Press, Boston.

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

Langacker, R. (1991). *Foundations of Cognitive Grammar. Practical Applications, Volume 2*. Stanford University Press, Stanford.

Mandler J (1999) Preverbal representations and language, in P. Bloom, MA Peterson, L Nadel and MF Garrett (Eds) *Language and Space*, MIT Press, 365-384

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

Morgan JL, Shi R, Allopenna P (1996) Perceptual bases of rudimentary grammatical categories, pp 263-286, in Morgan JL, Demuth K (Eds) *Signal to syntax*, Lawrence Erlbaum, Mahwah NJ, USA.

Quinn PC, Polly JL, Furer MJ, Dobson V, Nanter DB (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.

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

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

Siskind JM (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. *IEEE Intelligent Systems*, vol. 16, nr. 5, pp. 16-22, New York: IEEE Press.

Steels, L. and Baillie, JC. (2003). Shared Grounding of Event Descriptions by Autonomous Robots. Robotics and Autonomous Systems, 43(2-3):163--173. 2002

Stolcke A, Omohundro SM (1994) Inducing probablistic grammars by Bayseian model merging/ In *Grammatical Inference and Applications: Proc. 2nd Intl. Colloq. On Grammatical Inference*, Springer Verlag.

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