

A Developmental Robotic System for Visual Scene Perception and Language Acquisition

Corresponding Investigator:

Peter Ford Dominey, Ph.D. CNRS-CR1 Section 29 Telephone: 04 37 91 12 12
Institut des Sciences Cognitives Direct line: 04 37 91 12 66
CNRS UMR 5015 FAX: 04 37 91 12 10
67, Boulevard Pinel
69675 BRON Cedex FRANCE email: dominey@isc.cnrs.fr

Partners:

Peter F. Dominey *Institut des Sciences Cognitives, CNRS UMR 5015, Lyon*
Jean-Luc Schwartz, Christian Abry, *Institut de la Communication Parlée, Grenoble Cedex 1,*
Emmanuel Dupoux LSCP, Paris, France
Deb Roy *Cognitive Machines Group, MIT Media Laboratory, Cambridge, MA, USA*
Luc Steels, *AI Laboratory, Vrije Universiteit Brussel, Brussels Belgium*
Michael Arbib *Human Brain Project, University of Southern California*
Ram Nevatia *ISI, University of Southern California*
Laurent Itti, *Human Brain Project, University of Southern California*
Aude Billard, *Human Brain Project, University of Southern California*
Jeffrey Mark Siskind *Purdue University, Electrical and Computer Engineering, IN, USA*
M. Anthony Lewis, *Iguana Robotics, IL, USA*
Andrew H. Fagg, *University of Massachusetts Amhurst, Computer Science Dept, USA*

Situation analysis:

Humanoid robots exist and can be purchased for ~65K USD today. While their sensorimotor/postural control has been effectively resolved, robot cognition is in its infancy. Success in this field will be enhanced by coordinated cooperation between multiple international expert groups. Success will also require clear identification of long term and short term objectives, and a system architecture with well defined interfaces. A principal issue is the specification of an environment in which the robot is to operate. An initial proposal is that the robot should be seated across from a human in a setting that involves observation of the human manipulating objects, with the robot being capable of performance including: 1. Describing the events in an interactive discourse, 2. Mimicking or reproducing the events. A related objective is to maintain close contact with issues of developmental cognition, rather than to take a purely engineering approach.

Objectives of this document:

This document contains a "strawman" proposal for phase 1 of a humanoid robot cognition project. Its objective is to provide the point of departure for discussion for an international co-operative program that will allow joint funding at the national (e.g. NSF and CNRS for the US and France respectively) and international levels. Part of this funding would be for the purchase of equivalent robot platforms for the related institutions, and for the funding of graduate and post-graduate students, with the possibility of exchange between institutions.

Status:

This document is to be distributed in October of 2002. An initial planning meeting will take place in Lyon during January 28-29 2003, to coincide with a meeting on speech and localization organized by C. Abry in Grenoble. A 2-3 day workshop will then be organized with all participants in June 2-5 of 2003.

Proviso: This document should not be considered as a committed position statement, but rather as a work-in-progress document that is simply a starting point for discussion.

Abstract:

We have recently begun to address the issues of perceptual scene analysis and natural language interfaces (acquisition) from the perspective of developmental robotics. In the developmental trajectory of a human infant between 0-24 months of age both of these issues are addressed in a quite robust and effective manner. In particular, at 6 months of age, infants are able to “parse” complex and dynamic visual scenes to identify causal events and their agents and goals (Leslie & Keeble 1987, Woodward 1998), in a manner that is already equal or superior to current machine vision methods. Likewise, by 14 months of age, these infants have begun to construct the language-to-scene mapping capability that allows language and visual scene analysis to intersect in a common internal “conceptual scene” representation (Hirsh-Pasek & Golinkoff 1996). It is evident that there exists a highly productive and synergistic interaction between the processes of language acquisition, and visual scene analysis acquisition, that allow the infant to develop these two capabilities in a rapid and robust manner. Based on these observations, we have developed a functioning prototype of a “baby robot” that performs perceptual analysis of visual scenes, and constructs the mapping between natural language narration of scenes, and the internal representation of the analyzed scene.

The long term objective is to extend this developmental process to an autonomous real-time robot platform. The objective of the current project is to extend this prototype to allow a more robust treatment of visual scenes and natural language, including: the ability to processes multiple visual events and ongoing narrative in order to construct complex discourse-level scene representations, and the related ability to support an on-line interactive man-machine interface. The result will be an interactive autonomous system that can learn a natural language interface and then use this interface in dialogs about the ongoing state of perceived complex dynamic visual events. We have already demonstrated a significant capability in this direction, based on the development of our existing operational prototype over the last 12 months. Our teams are well adapted to the proposed objectives, and the project can thus be considered to be highly feasible.

Abstract:.....	2
Overview:.....	3
Existing Prototype System and Underlying Hypotheses:	3
Operation of the Prototype:	4
1. Speech Input processing (Human generated speech).....	4
2. Lexical Analysis (Lexical_analysis.C)	4
3. Visual Scene Input (Human generated actions)	4
4. Low Level Vision (Panlab SMART Video Tracking System).....	5
5. Visual Scene Analysis (Scene_analysis.C).....	5
6. Language processing (LanguageAcquisition.C).....	6
Results and Lessons Learned from the Prototype:	7
Scientific Objectives:	8
1. Increased Contextual Complexity of Events	8
2. Increased Interaction:	8
3. Testing and Portability and Robustness for future use with humanoid robots.	9
Methodological Approach:.....	9
1. Increased Contextual Complexity of Events.....	9
2. Increased Interaction:	10
3. Testing and Portability for future use with humanoid robots.	11
Project Planning (suggestive – not to be interpreted literally)	15

Overview:

In the current call for proposals, four of the five top priorities are (1) perception for scene interpretation, (2) autonomous agent interaction via multimodal communication including natural language, (3) learning and extraction of structure from the environment, and (4) man-machine interaction. We have recently begun to address these issues from the perspective of developmental robotics. Specifically, we have developed a functioning prototype of a “baby robot” that performs perceptual analysis of visual scenes, and constructs the mapping between natural language narration of scenes, and the internal representation of the analyzed scene based on the functional architecture in Figure 1.

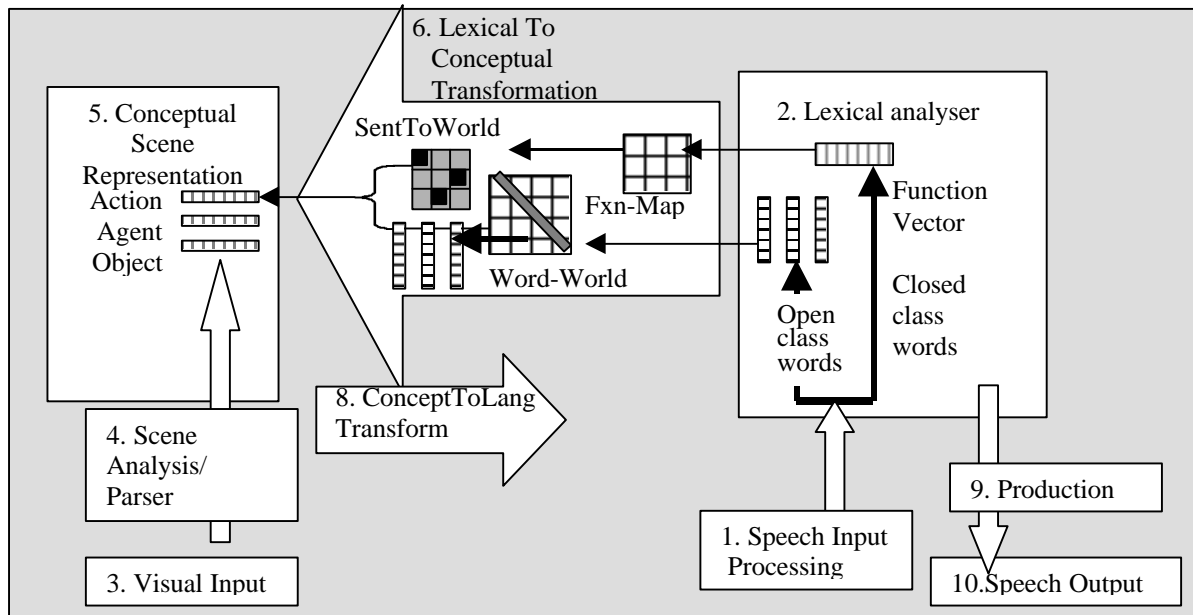


Figure 1. Language and Vision Processing Architecture. Input sentence dissociated into open class (nouns, verbs, adjectives) and closed class (grammatical function) words. Open class word meaning is retrieved (Word-World) and thematic roles are assigned (SentToWorld) based on syntactic structure (FunctionVector and Fxn-Map). This information feeds into the conceptual scene representation, that also receives input from the visual scene analyzer/parser. The combined visual and linguistic inputs permit word and syntax learning, scene-sentence comparison, verbal completion of missing visual information, etc.

Existing Prototype System and Underlying Hypotheses:

The prototype demonstrates that a robotic system that perceives visual scenes and speech can use associative learning to construct a mapping between structure in visual scenes, and the grammatical structure of sentences that describe those scenes. This learned mapping allows the system to process natural language sentences in order to reconstruct complex internal representation of the visual scenes that those sentences describe. In the system, low level perceptual processes of speech segmentation and visual object recognition and tracking are provided by commercial software products. Structure is then further extracted from these representations, and a novel associative learning technique is employed to establish the mappings between different grammatical structures and the corresponding event structure of the paired visual scene. During post-learning sentence interpretation, the appropriate mapping of grammatical structure to scene structure is retrieved based on grammatical markers inherent to the sentence. The system demonstrates error free performance for a rich subset of English that includes complex hierarchical grammatical structure.

Operation of the Prototype:

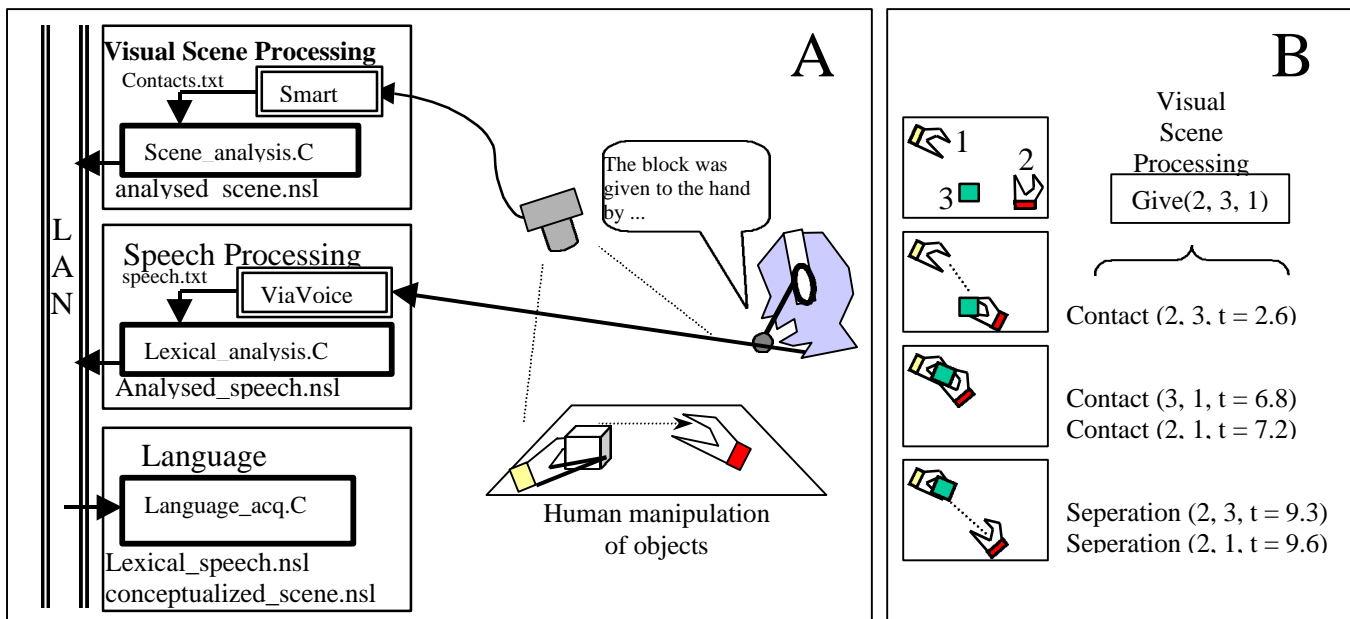


Figure 2. Implemented prototype. Allocation of processing functions to hardware, with indication of processing and data-flow.

1. Speech Input processing (Human generated speech)

Input: Human speech input is processed in real time during human action narration. The speech input is processed by a commercial/public domain speech recognition software (IBM ViaVoice Pro, V8, American English).

Output: The output is an ASCII text representation of the processed speech. (Speech.txt)

2. Lexical Analysis (Lexical_analysis.C)

Input: Text input from the “Speech input processing” component (speech.txt).

Processing: *Simplification 1:* Words in the speech stream are segregated into two broad categories: open class (nouns, verbs, adjectives, etc.) and closed class or function words (determiners, prepositions, etc.). This is supported by extensive behavioral and neurophysiological data. (Dominey – Lexical_Analysis.C)

Output: (Analysed_speech.nsl) Words are recoded single bits in a 25 element vector, with open and closed class words coded in distinct subregions of the vector. This will lead to dissociated processing of closed class words in the Function vector (a representation of the closed class and functional projections), and open class words in the Open Class Array (an ordered set of open class elements) in LanguageAcquisition.C. Note that in the prototype, with this coding scheme, the initial vocabulary is limited to 25 words.

3. Visual Scene Input (Human generated actions)

Visual input is provided from a CCD camera in real-time,. The visual scenes consist of human manipulation of simple objects, enacting “touch, push, give, take” and related actions, described in more detail below.

4. Low Level Vision (Panlab SMART Video Tracking System)

Input: Visual Scene Input (3)

Processing: *Simplification 2* : Color-based object recognition: The SMART system tracks multiple objects based on their dissociable colors. The system tracks absolute, and relative position and velocity between objects.

Output: (contacts.txt) that contains:

- 1) A time ordered list of contact events, i.e. descriptions of which objects have come within a minimal distance (parameterized) of each other.
- 2) Discrete timed (5Hz) description of all objects, their locations, relative positions and velocities for all object pairs.

5. Visual Scene Analysis (Scene_analysis.C)

Input: contacts.txt from Low Level Vision.

Processing: Construct a higher level representation of scene events, in the form of a list of events and their arguments.

Simplification 3: Scene events are made up of contacts. A contact between two physical elements 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. Interestingly, this parameter of movement is also one of the most perceptually salient visuo-spatial properties used by human infants in scene analysis. Based on these parameters of contact, scene events are recognized as follows:

- Touch(agent, object) : This event corresponds to 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) : This event corresponds to a single contact in which (a) the duration of the contact is superior or equal to *touch_duration* and inferior to *take_duration* (5 sec), (b) the object is displaced during the duration of the contact, and (c) the agent and object are not in contact at the end of the event.
- Take(agent, object) : This event corresponds to 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) : In this event, the agent takes the object from the source. This is a compound event that is identical to Take(agent, object) with a second contact between the agent and the source in which (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) : In this event, the agent gives the object to the recipient. This is a compound event, made up of multiple contacts in which the agent gives the object to the recipient. For the first contact between the agent and the object (a) the duration of contact is inferior to *take_duration*, (b) the object is displaced during the contact, and (c) the agent and object do not remain in contact. For the second contact between the object and the recipient (a) the duration of the contact is superior to *take_duration*, and (b) the object and recipient remain in contact. For the third (optional) contact between the agent and the recipient (a) the duration of the contact is inferior to *take_duration* and thus the elements do not remain in contact.

Complex Events:

The events described above are simple in the sense that there have no recursive or hierarchical component. This imposes serious limitations on the syntactic complexity of the corresponding sentences (Feldman et al. 1996). The sentence “The block that pushed the

moon was touched by the triangle” describes a complex event that exemplifies this issue. Such a compound event will be recognized and represented as a pair of 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.

These event labeling templates form the basis for a template matching algorithm that labels events based on the contact list.

Output: The output of this “high-level” vision is the file *analysed_scene.nsl*, a time ordered list of event descriptions.

6. Language processing (LanguageAcquisition.C)

This is the language acquisition model illustrated in Figure 1. Open class lexical items will be associated with their world/scene counterparts. Based on the configuration of functional elements within the sentence, a mapping will be constructed from relative sentential position to functional (thematic) role.

Inputs:

1. *analysed_speech.nsl* – contains information assigning words to Function vector and Open class array from the Lexical analyzer.
2. *analysed_scene.nsl* - Analyzed Scene from the Scene Analysis.

Processing:

1. *Mapping open class words to their conceptual representation:* For each lexical element of the Open Class Array that element will be transformed into its corresponding conceptual representation by the matrix of modifiable (learned) connections *WordToWorld*.

2. *Mapping grammatical structures onto event structures : Simplifying Hypothesis :* For every grammatical sentence in the language, there is an unambiguous mapping between the open class elements in the sentence, and their corresponding elements in the scene event representation. These mappings will be represented in the *SentenceToWorld* matrix. *Simplifying Hypothesis :* Each grammatical sentence category or grammatical form (examples in Table 1B) can be uniquely represented by a “FunctionVector” constructed from a concatenation of the function words in that sentence. The function vector will then serve as an index into an associative memory to retrieve the appropriate *SentenceToWorld* matrix for the given grammatical form, as $SentenceToWorld = FunctionVector \times FxnMap$. *FxnMap* is a matrix of modifiable connections that is established by learning. This retrieved matrix specifies the mapping of open class elements onto their thematic/conceptual roles (Note that the matrix *SentenceToWorld* is represented above as a vector to simplify the multiplication notation.)

Learning: Thus, the initial steps of language acquisition will be in establishing the “word-to-world” and then “sentence-to-world” mappings.

1. *WordToWorld:* The mapping from open class lexical items onto their world/scene counterparts is learned by association. The learning algorithm initially associates all open class words in a sentence, with all elements in the conceptualized scene, exploiting cross-situational statistical regularities (Siskind 1996). Later, this learning is refined by using syntactic knowledge (in *SentenceToWorld*) in order to only associate lexical items with the syntactically corresponding element of the conceptual scene (syntactic bootstrapping).

2. *SentenceToWorld:* Each grammatical form generated by the language (e.g. “The ball was pushed by the block”) will become associated to a specific mapping between the open class array and the conceptual scene array (e.g. “push(block, ball)). These mappings are encoded in the *SentenceToWorld* array for the different grammatical forms (see Table 1B). As each sentence is processed, the estimated *SentenceToWorld* mapping is generated, based on existing *WordToWorld* correspondences. Learning then consists in building the association between

the current FunctionVector, and this estimated SentenceToWorld mapping (for the current grammatical form), stored in the FxnMap.

Results and Lessons Learned from the Prototype:

The existing prototype can learn a reduced version of the English language (characterized by the context free grammar in Table 1A below). That is, it learns the mapping of multiple grammatical forms - based on active, passive and relative structures - (depicted in 1-10 of Table 1B below) onto the corresponding representations of the paired visual scenes. The system can subsequently determine if a given sentence corresponds (or not) to a given visual scene. Indeed the system is remarkably robust in the ability to learn the structural mapping from grammar to scene, as illustrated in Figure 4.

<ol style="list-style-type: none"> 1. S → NP + VP 2. NP → Det + N 3. NP → NP + Rel + VP 4. VP → Va + NP 5. VP → Va + NP + PP 6. VP → Aux + Vp + PP 7. VP → Aux + Vp + PP + PP 8. PP → Prep + NP 9. Det → the, a 10. N → cylinder, block, moon 11. Va → touched, pushed, took, gave 12. Vp → touched, pushed, taken, given 13. Aux → was 14. Prep → to, by, from 14. Rel → that 	<ol style="list-style-type: none"> 1. Active: The block pushed the triangle. 2. Dative: The block gave the triangle to the moon. 3. Passive: The triangle was pushed by the block. 4. Dat Pass: The moon was given to the triangle by the block. 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.
<p>A. The grammar</p>	<p>B. Sentence types (grammatical forms) generated by the grammar</p>

Table 1. A. The context free grammar corresponding to the language that is learned by the system. B. A set of grammatical forms that are generated by the grammar.

The results in Figure 4 illustrate that like human infants, in the totally naive state (Exp A) the model must learn both word meanings (stored in the Word-World matrix Fig. 1), and sentence grammatical structure. Indeed, we have demonstrated that these two processes must interact in a synergistic manner, or the system can never resolve the semantic and syntactic ambiguities. Once the word meanings have been acquired, the system can then develop a more robust and complex representation of grammatical structure, which, in turn, can be exploited when new words must be learned in these grammatical forms (Dominey 2002).

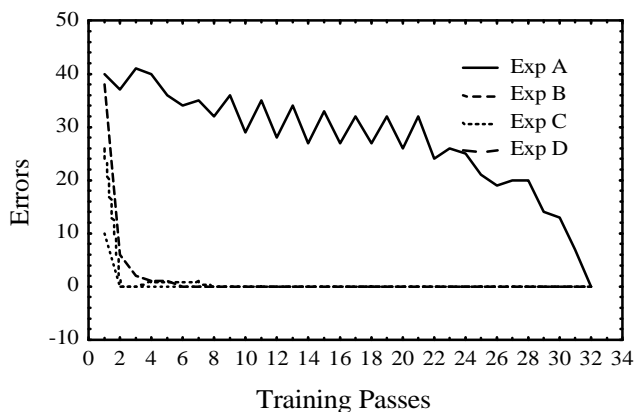


Figure 4. Performance – Number of interpretation errors during successive training passes through the set of 27 sentences. Exp A. Sentence types 1 and 2 (from Table 1B), and word meanings are learned in parallel. Exp B. Sentence types 3 and 4 learned with fixed word meanings (acquired in A). Exp C. Sentence types 5-10. Exp D. Sentence types 1-10. Note that once word meanings have been acquired, mapping grammatical to scene structure is highly effective.

Limitations: Despite these promising results, there are, however, several weaknesses that have been identified and that must be improved. The capability to represent context and complex temporal events is limited. The verbal input is currently restricted to single sentences that correspond to simple one or two verb events, thus, complex scenes that evolve in time cannot currently be represented. The current system does not produce language/speech output that would allow man-machine interaction. The current system has a fixed visual system that is sensitive to lighting conditions, and thus is not well adapted for a mobile platform. The current system operates in batch mode, without interactive processing of single scene/sentence pairs.

Scientific Objectives:

The long-term objective is to extend the prototype to provide the visual and verbal interface to a fully interactive mobile humanoid robot (e.g. Fujitsu, Honda, Sony). Part of our approach is to use off-the-shelf technology where available, so that we can concentrate on the technical issues of scene analysis and the natural language interface. In order to achieve this goal, and to address the limitations identified by the prototype study, we have identified 3 principal objectives to be achieved during this two-year project:

1. Increased Contextual Complexity of Events.

The first objective is to allow perception and interpretation for multiple-event representations in the context of previous events, based on visual scene analysis and natural language input.

1.a. Visual scenes (and the corresponding events) are rarely meaningful in isolation from the dynamic context. They tend to occur in an ongoing temporal frame. Thus, a given scene event should be interpreted in the context of previous events, and possibly in the context of future events.

1.b. Likewise, a single sentence is rarely meaningful (e.g. "Put it on the table." What does "it" refer to?). Sentences tend to occur in an ongoing discourse and thus should be interpreted in a discourse context.

With respect to any interaction between man and machine, this objective is highly important, as any meaningful interaction will involve multiple related sentences that describe some complex ongoing event.

2. Increased Interaction:

The second objective is to allow interactive processing of speech and vision, in order to allow the interactive dialogue between the system and the human that will be required for the first objective. In the current implementation, speech and vision data are gathered in a batch mode, during ~5 minutes corresponding to approximately 25 separate visual scene events, and the corresponding separate sentences.

2.a. Visual scene analysis. Bioseb/Panlab Smart Vision currently requires the experimenter to initiate and then terminate the visual scene capture (multiple subjects tracking), and then to exit the tracking program, enter the analysis program, specify the file etc. in order to generate the ASCII file of contact events. The objective will thus be to employ a visual scene analysis with near real-time automatic generation of the ASCII output file for single or multiple events based on pre-specified triggers.

2.b. Speech and language processing: IBM ViaVoice is used to generate ASCII text that can be accessed on-line. The current implementation of the language acquisition/processing model operates in batch mode. The objective is to allow for on-line processing, as well as the possibility of the system to interrogate the human operator.

3. Testing and Portability and Robustness for future use with humanoid robots.

In addition, we frame these objectives in the longer-term context of their implantation in an autonomous mobile (humanoid) robotic platform with object manipulation capability. This will also involve validation of the proposed system on humanoid torso and locomotion platforms.

Methodological Approach:

1. Increased Contextual Complexity of Events.

The Prototype system used relatively simple events, and a relatively simple visual system. These issues are addressed as follows.

1.1 **Increased Event Complexity** We will now introduce goal based actions. For each of these actions, we will define 2-D and 3-D temporal geometry models that will allow the scene analyzer to recognize these actions (see 1.2.2).

1.1.1 Put the blocks in the box: A sequence of “take” events in which the initial location of the objects is in the box, and the final location is outside of the box

1.1.2 Get the blocks out of the box: Similar to 1.1.1 in the opposite sense.

1.1.3 Build an arch: A sequence of object placements that results in an arch as illustrated in Figure 3. Note the importance of camera angle.

1.1.4 Build a pyramid

1.1.5 Build a house

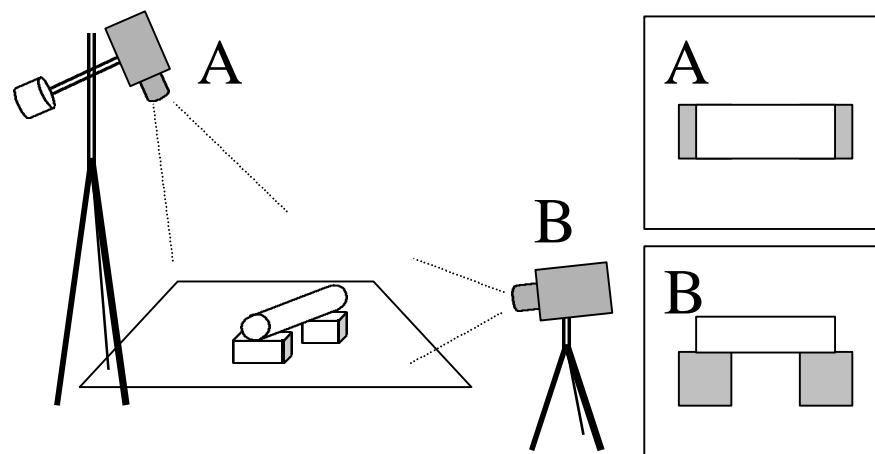


Figure 3. Camera angles and resulting views of an “arch”.

1.2 **Improvement of Visual Processing.** Visual processing consists of (1.2.1) object recognition and tracking, and (1.2.2) event extraction, that will be extended to 3-D model construction.

1.2.1 3-D Recognition and Tracking: Our current color-based recognition and tracking method (Smart) is relatively robust in our controlled experimental conditions, but will not extend well to more unconstrained environments. Thus we will initially continue to use Smart for recognition and tracking, and introduce more robust recognition and tracking as they are developed.

1.2.2 Complex Event Extraction: In parallel we will begin to integrate simplified 2-D and 3-D geometric models of these events outlined above (e.g. Fig 3B) into the event extraction

component of visual processing. For each of the complex events in 1.1 we will define a temporal event logic description or model (Siskind 1997, 2000) that will allow automatic detection of these events. This will thus extend the existing event detection capability of the prototype based on contact, to consideration of contact, support and attachment (Siskind 1997, 2000). Initially the 2-D models will be pre-defined, though we will also investigate the possibility of learning new model descriptions from the environment, based on Siskind (1996b).

1.3 Modification of the scene representation. Figure 1 indicates a highly simplified scene description in terms of a single action and its agent and object. We will modify this representation to become a dynamic frame based representation (corresponding to structures in 1.1) that can continue to be accessed by the perceptual analyzer and by natural language input.

1.4 Vision/Internal representation Interaction: The scene interpretation algorithm is currently feed forward. In the new implementation, the scene interpretation algorithm will have access to the ongoing construction of the conceptual scene representation, which will provide additional context for disambiguation.

1.5 Language/Internal-representation interaction: Similarly, the structure mapping component of the language processor will also have access to the dynamic scene representation in order to resolve referential ambiguities.

2. Increased Interaction:

An interactive man-machine dialog will require the use of a dialog management system. We will use the MIT/Mitre/CMU Galaxy Communicator architecture as a model. In this context, the system will be extended in the following manner. Note that Items 2.1 and 2.2 are the specific priority goals and 2.3, 2.4 and 2.5 are included for completion if possible.

2.1 Understand Vision and Language: The prototype is currently able to “understand” in the sense that it can construct the mapping between sentence structure and visual scene structure. This will be extended to the increased complexity events described above. For example, the Arch, House, and Pyramid models will specify the relative geometric configurations of the component elements, and temporal constraints on how the elements are to be placed. The representations will thus be of a frame-based nature, extending those described in the prototype section.

2.2 Describe visual scenes: In Fig. 1 we see that the scene representation can be accessed both by the vision system and the language system. The proposed system shall thus be able to access the scene representations in order to drive the language stream in the “reverse” sense for sentence production, in either active or passive voices depending on the context.

2.3 Respond to Questions: In a related sense, the system shall be capable of responding to questions concerning the current scene/situation. Questions will be processed to determine the focus (the event itself, an element of the event), and then this focus will be used to interrogate the event representation.

2.4 Interrogate: Finally the system should be capable of interrogating the human to resolve processing ambiguities. Questions generation will be triggered by the detection of missing information, using standard “wh” transformations.

2.5 Increased vocabulary. The current vocabulary size is limited to 25 because of the single bit coding. Siskind (1996, 2000) has developed efficient and robust cross-situational learning methods that will allow us to increase this into the 10^3 word range.

2.6 Technical issues of interactive processing:

2.6.1 Visual scene analysis. The Bioseb/Panlab Smart Vision, and the final 3-D system will allow near real-time automatic generation of the ASCII output file for single or multiple events based on pre-specified triggers.

2.6.2 Speech and language processing: IBM ViaVoice will be used to generate ASCII text that can be accessed on-line. The language acquisition/processing model will be modified so that processing of a given input sentence (or sentences) will be initiated in a message driven protocol in the context of the dialog manager.

3. Testing and Portability for future use with humanoid robots.

The long term objective is that this grammatical-to-perceptual structure mapping capability can be integrated into one or more humanoid robot control systems. This will be accomplished in two manners.

3.1 Development of a detailed requirements document that defines how the structure mapping system will interface and operate with the following systems

3.1.1 The UMASS Humanoid Torso (Andrew Fagg – Chief Architect).

3.1.2 The Iguana Robotics walking robotic platform (Tony Lewis – Director)

3.1.3 Commercial humanoid robot architectures such as the Fujitsu Human Robotics Program Prototype 2 (HRP-2P), on ART-Linux

3.2 Based on the requirements defined in 3.1.1 the grammatical-to-perceptual structure mapping capability will be tested on the UMASS Humanoid Torso. The UMass Torso project focuses on questions of human development as embodied in a robotic system. This system is roughly anthropomorphic and consists of a pair of 7 degree-of-freedom Whole Arm Manipulators (WAMs), a pair of three-finger hands, and a binocular stereo head. Each finger tip is equipped with a force-torque sensor that supports the generation of haptically guided grasping behavior. Tasks are performed through the sequential activation of sets of concurrent controllers that satisfy contact placement/force and kinematic conditioning constraints (Coelho and Grupen, 1997). System state is represented through controller convergence events and the dynamics of interaction between the controllers and the objects being grasped.

The proposed techniques for establishing relationships between linguistic inputs and observed visual events can also be employed to construct similar representations between linguistic data and both haptic inputs and sequences of controller activations. Initial training of such representations can be accomplished through the human narration of the actions taken by the robot as it proceeds through a set of reaching and grasping tasks. Once established, these representations may be used to instruct the robot through a novel sequence of actions. Furthermore, it becomes possible for the human and robot to perform cooperative tasks, including the human making requests such as "take the blue block from me," to the robot making requests including "please fetch the purple cylinder for me."

3.3 Based on the requirements defined in 3.1.2 the grammatical-to-perceptual structure mapping capability will be tested on the Iguana walking platform. The RoboKid Project is directly complementary to this existing research project. While both project use a developmental strategy, the Iguana Humanoid project is emphasis learning 'low level' visuomotor tasks related to walking. The RoboKid Project, building on a similar strategy, could add a dimension of human interactivity that is missing from Iguana current project. Conversely, the RoboKid project does not address low-level visuomotor control of gait walking will benefit from the being hosted on a highly mobile platform suitable for real-world circumstances.

Ultimately, the RoboKid project can become the controlling interface for the Iguana Humanoid Robot. It will be important from the outset to maintain close ties with the RoboKid group in order to facilitate a smooth integration of these complementary systems.

Criteria for Determining Success: Here we provide two concrete processing examples. In the first example, the system observes an event and is then asked a question.

Events	Processing
Scene Event 1: block 1 pushes block 2	<i>Push(block1, block2)</i> event detected [1]
Narration Event 1: Speaker asks “What happened to block 2?”	Interrogation with “block 2” as the narrative focus.
System responds: “Block 2 was pushed by block1.”	Retrieve grammatical form for 2 argument verb with the object role in the focus (first word) position, and apply this form to the scene event <i>Push(block1, block2)</i> [2]

[1] As in the prototype.

[2] This implies (1) that when the forward transformation from grammar to scene is learned, the reverse transformation from scene to grammar shall also be learned, (2) that grammatical forms can be indexed according to these criteria (number of arguments, and discourse focus).

Here we provide a concrete example of a scenario in which the system observes and then describes a “build an arch” event:

Events	Processing
Scene Event 1: block 1 is taken from the holding area	Block 1 changes state, and its new position enters geometric model [1]
Scene Event 2: block 2 is taken from the holding area and placed near block 1	Block 2 position, and “nearness” relation between blocks 1 and 2 enters geometric model [2]
Scene Event 3: block 3 is placed over blocks 1 and 2 as indicated in Fig B.	Block 3 position, and “arch” relation between blocks 1,2 and 3 enters geometric model, as a complex 3-D construction “ arch(left(block1), right(block2), top(block3)) ” [3]
Narration Event 1: Speaker asks “What is there?”	Speech system decodes “what” question, and uses the 3-D geometry description “arch(left(block1), right(block2), top(block3))” in speech production mode
System responds: “There is an arch made of block1, block2 and block3.”	
Narration Event 2; Speaker asks “What is block3 doing”	Speech system decodes “what” question, and uses the state event “arch(left(block1), right(block2), top(block3))” with reference to the role of “block three” in speech production mode
System responds: “Block3 is on top of the arch”	

[1] The system is “reactive” in that it “pays attention” to state changes.

[2] This includes contact and proximity states

[3] During previous training, this arch state-event has been paired with utterances of the form “There is an arch made of block1, block2 and block3,” similar to the way the current prototype has been exposed to simple events and their narration.

Deliverables: Here we clearly identify what will actually be produced or “delivered”.

1. RoboKid System (Lyon).

The first deliverable is the system itself, consisting of an ensemble of hardware and software.

1.1 The system shall be capable of:

- 1.1.1 Processing visual and verbal descriptions of complex multiple-action events
- 1.1.2 Production of verbal (audio) description of events
- 1.1.3 Capability to respond to event-related questions.

- 1.2 The system shall be compatible with humanoid robot requirements for scene processing and motor control, as specified in 2 below. In particular, the system shall be tested on two distinct platforms :
 - 1.2.1 The UMASS Humanoid torso
 - 1.2.2 The Iguana Locomotion Platform
2. System Requirements Document (SRD) for Humanoid Robot Perception and Cognition (HRPC).
 - 2.1 The second deliverable shall be a requirements document for an extension of the RoboKid system to a complete sensory-motor-cognitive control system that is compatible with installation on an identified humanoid robot platform. The SRC-HRPC shall identify requirements on:
 - 2.1.1 General requirements for a cognitive system for humanoid robots.
 - 2.1.2 Vision
 - 2.1.3 Language
 - 2.1.4 Prehension/Manipulation
 - 2.1.5 Locomotion
 - 2.1.6 Representation
 - 2.2 The SRC-HRPC shall specify the shared data-structures that define the interface between the grammar-to-perception mapping algorithm, and the perceptual representations used by
 - 2.2.1 The UMASS Humanoid torso
 - 2.2.2 The Iguana Locomotion Platform
 - 2.2.3 Existing humanoid robot platforms such as the Fujitsu HRP-2P

Resources (partial list – to be completed):

Dominey:

Related Funding: The development of the existing prototype at the ISC in Lyon/Bron has been supported in part by a grant from the French Minister of Research under the ACI Neuroscience Integrative et Computational (ACI-NIC) for the project: *Adaptation fonctionnelle pour l'acquisition de la parole et du langage chez un «bébé robot»*. Participants/consultants in that project include Jean-Luc Schwartz, Luc Steels, Emmanuel Dupoux, Michael Arbib, Ram Nevatia and Deb Roy. The (ACI-NIC) funding is to support equipment and laboratory operating expenses for the prototype system.

Our group has also received funding from the project: Contrat “COGNITIQUE”, 1999-2001 sur Thème 2 “Perturbations et récupération des fonctions cognitives” : *A Neural Network Model of Syntactic Comprehension and its Application to Rehabilitation in Agrammatic Aphasia*. This contract provided support for the refinement of the language acquisition model, and for the execution of a series of human neurophysiology experiments to validate the hypotheses underlying the model.

Resources available: We have the current prototype described above, that consists of two 1GHz PCs for vision (Smart) and speech (ViaVoice) processing with a Sun Unix and PC Linux workstations for language processing and software development.

Siskind:

Related funding: No specific funding for this project.

Resources available: Software packages for event classification based on motion, and force dynamics.

Lewis:

Related funding: Dr. M. Anthony Lewis is the President of Iguana Robotics that leads a consortium of four research institutions in the construction of an intelligent biped robot based on biological principles, under the sponsorship of the Office of Naval Research (ONR). The project's approach is to perform human experiments to determine how humans modify gait trajectories based on visual input. This data will then be modeled and hosted in a special purpose neural computer, controlling a walking robot. Extensive use of developmental learning strategies will be used. The final goal of the project is to create a humanoid capable of rapidly moving through its environment. The device will learn to move more accurately, at higher speeds and with more coordination through the use of novel learning algorithms.

Resources available: Iguana Humanoid Robot platform. The RoboKid Project is directly complementary to this existing research project. While both project use a developmental strategy, the Iguana Humanoid project is emphasis learning 'low level' visuomotor tasks related to walking. The RoboKid Project, building on a similar strategy, could add a dimension of human interactivity that is missing from Iguana current project. Conversely, the RoboKid project does not address low-level visuomotor control of gait walking will benefit from the being hosted on a highly mobile platform suitable for real-world circumstances.

Ultimately, the RoboKid project can become the controlling interface for the Iguana Humanoid Robot. It will be important from the outset to maintain close ties with the RoboKid group in order to facilitate a smooth integration of these complementary systems.

Fagg:

Related funding: Andrew Fagg is currently supported under a Research Infrastructure grant from the National Science Foundation (Prof. Rod Grupen, PI). This work focuses on understanding aspects of the development of reaching/grasping skills in human infants and implementing models of these developmental processes on a humanoid robot. This robot (the UMass Torso) consists of 2 seven degree-of-freedom arms, 2 three-fingered hands, and an articulated stereo camera system..

Continued support is pending from the National Science Foundation and the Defense Advanced Research Projects Agency (DARPA) (Grupen/Fagg PIs), and will focus on learning of reaching/grasping/manipulation skills and how these skills are then applied in the context of solving tasks.

Resources available: The UMass Torso consists of a pair of 7 degree-of-freedom Whole Arm Manipulators (WAMs), a pair of three-finger hands, and a binocular stereo head. Each finger tip is equipped with a force-torque sensor that supports the generation of haptically guided grasping behavior.

Project Planning (suggestive – not to be interpreted literally)

	Sept-Dec 2002	Jan-June 2003	Sept-Dec 2003	Jan-June 2004
1. Increased Event Complexity				
1.1 Complex 3-D events	x			
1.2 Improved vision				
1.2.1 3-D Recog & Track (Rennes)	x	x	x	x
1.2.2 Complex Event Extraction (Lyon)	x	x		
1.3 Contextual Scene Representation	x			
1.4 Vis-Scene Interaction		x		
1.5 Lang/scene Interaction		x		
2. Increased Interaction				
2.1 Vis/Lang complex events	x			
2.2 Scene description		x		
2.3 Respond to questions			x	
2.4 Interrogation			x	
2.5 Vocabulary		x	x	
3. Portability and Testing				
3.1 Requirements Document	x	x		
3.2 Test UMass			x	x
3.3 Test Iguana			x	x
Workshops		x		x

Partial References:

- Abry, C., Boë, L.J., Laboissière, R., & Schwartz, J.L. (1998). A new puzzle for the evolution of speech ? *Behavioral and Brain Sciences*, 21, 512-513.
- Arbib, M.A., 1997, Modeling visuomotor transformations, in Handbook of Neuropsychology, Volume 11, Section 16: Action and Cognition, (M. Jeannerod, Ed.), Amsterdam: Elsevier, pp.65-90.
- Arbib, M.A., Billard, A., Iacononi, M., and Oztop, E., 2000, Synthetic Brain Imaging: Grasping, Mirror Neurons and Imitation, *Neural Networks*, 13: 975-997.
- Billard, A. and Dautenhahn K., 1998, Grounding communication in autonomous robots. *Robotics and Autonomous Systems*, Vol. 24, Issues 1-2, p. 71-79, 1998.
- Christophe, A., Guasti, M.-T., Nespore, M., Dupoux, E., & van Ooyen, B. (1997). Reflections on phonological bootstrapping: its role for lexical and syntactic acquisition. *Language and Cognitive Processes*, 12, 585-612.
- Dehaene, S., Dupoux, E., Mehler, J., Cohen, L., Paulesu, E., Perani, D., van de Moortele, P.-F., Léhericy, S., & LeBihan, D. (1997). Anatomical variability in the cortical representation of first and second languages. *Neuroreport*, 8, 3809-3815.
- Dehaene-Lambertz, G., Dupoux, E., & Gout, A. (2000). Electrophysiological correlates of phonological processing: a cross-linguistic study. *Journal of Cognitive Neuroscience*, 12, 635-647.
- Dominey PF (2002) Conceptual Grounding in Simulation Studies of Language Acquisition, (In press) *Evolution of Communication*
- Dominey PF (2002) Neurological basis of language in sequential cognition: Evidence from simulation, aphasia and erp studies, (In press) *Brain and Language*
- Dominey PF, Ramus F (2000) Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant. *Language and Cognitive Processes*, 15(1) 87-127
- Dupoux, E., Pallier, C., Kakehi, K., & Mehler, J. (in press). New evidence for prelexical phonological processing in word recognition. *Language and Cognitive Processes*
- Feldman J., G. Lakoff, D. Bailey, S. Narayanan, T. Regier, A. Stolcke (1996). L0: The First Five Years. *Artificial Intelligence Review*, v10 103-129.
- Garcia L-M, A. A.F. Oliveira, R.A. Grupen, D. S. Wheeler, A. H. Fagg (2000) Tracing Patterns and Attention: Humanoid Robot Cognition, *IEEE Intelligent Systems*, 15(4) 70-77.
- Hirsh-Pasek, K., Golinkoff R.M (1996) "The origins of Grammar", MIT Press
- Hongeng S., Bremond F., Nevatia R. (2000) Representation and Optimal Recognition of Human Activities. In *IEEE Proceedings of Computer Vision and Pattern Recognition*, South Carolina, USA,.
- Itti L., C. Koch, E. Niebur, A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 20, No. 11, pp. 1254-1259, Nov 1998
- Itti, L., and Koch, C., 2001, Computational Modelling Of Visual Attention, *Nature Reviews Neuroscience* 2:194-203
- Leslie AM, Keeble S (1987) Do six-month-olds percieve causality ? *Cognition* 25, 265-288.
- Lewis MA, (2002) "Gait Adaptation in a Quadruped Robot", *Autonomous Robots*, in Press
- Lewis MA, Kar-Han Tan (1997) High Precision Formation Control of Mobile Robots using Virtual Structures, *Autonomous Robots*,
- Platt, Jr., R., Fagg, A. H., Grupen, R. A. (2002), Nullspace Composition of Control Laws for Grasping, Accepted, *International Conference on Intelligent Robots and Systems (IROS'02)*

- Roy D. & A. Pentland. Learning words from sights and sounds: A computational model. *Cognitive Science*, pp.
- Roy, D. Grounded Speech Communication. Proceedings of the International Conference on Spoken Language Processing, 2000.
- Roy, D. Integration of Speech and Vision using Mutual Information. Int. Conf. Acoustics, Speech and Signal Processing, 2000.
- Roy, D. Grounded Language Acquisition: Experiments in Word Learning. In review, IEEE Transactions on Multimedia.
- Schwartz, J.L., Abry, C., Boë, L.J., & Cathiard, M. (2000). Phonology in a theory of perception-for-action-control. In J. Durand, B. Laks (eds.) *Phonology : from Phonetics to Cognition*. Oxford: Oxford University Press (in press).
- Schwartz, J.L., Robert-Ribes, J., & Escudier, P. (1998). Ten years after Summerfield ... a taxonomy of models for audiovisual fusion in speech perception. In R. Campbell, B. Dodd & D. Burnham (eds.) *Hearing by eye, II. Perspectives and directions in research on audiovisual aspects of language processing* (pp. 85-108). Hove (UK) : Psychology Press.
- Siskind JM (1996) A computational Study of Cross-situational Techniques for Learning Word-to-Meaning Mappings, *Cognition*, 61, 39-91.
- Siskind JM (1996b) Unsupervised Learning of Visually-Observed Events. In AAAI Fall Symposium Workshop on Learning Complex Behaviors in Adaptive Intelligent Systems, pp. 82-83
- Siskind JM (1997) Visual event perception, IN Ikeuchi & Veloso (Eds) *Symbolic Visual Learning*, chap. 9, NY, Oxford University Press.
- Siskind JM (2001) Grounding the Lexical Semantics of Verbs in Visual Perception Using Force Dynamics and Event Logic, *Journal of Artificial Intelligence Research*, volume 15, pp. 31-90,
- Steels, L. (1998) The Origins of Syntax in visually grounded robotic agents. *Artificial Intelligence* 103, 1-24.
- Steels, L. (2000) The emergence of Grammar in Communicating Autonomous Robotic Agents. In: Horn, W. (ed.) *Proceedings of ECAI 2000*. IOS Publishing, Amsterdam.
- Steels, L. and Kaplan, F. (1998) Situated Grounded Word Semantics. In *Proceedings of IJCAI-99*, Stockholm. Morgan Kauffman Publishing, Los Angeles. p. 862-867.
- Steels, L. F. Kaplan, A. McIntyre, J. Van Looveren (2001) Crucial factors in the origins of word meaning. In: *Proceedings of the 3d conference on the evolution of language*. Paris 2000.
- Wheeler, D. S., Fagg, A. H., Grupen, R. A. (2002), Learning Prospective Pick and Place Behavior, *Proceedings of the International Conference on Development and Learning (ICDL'02)*
- Woodward AL (1998) Infants selectively encode the goal object of an actor's reach. *Cognition* 69 1-34.