

# Building an Adaptive Spoken Language Interface for Perceptually Grounded Human-Robot Interaction

Peter Ford Dominey<sup>1</sup>, Jean-David Boucher<sup>1</sup>, Toshio Inui<sup>2</sup>

<sup>1</sup> Institut des Sciences Cognitives, CNRS  
67 Blvd. Pinel, 69675 Bron Cedex, France  
<dominey, boucher>@isc.cnrs.fr  
<http://www.isc.cnrs.fr/dom/dommenu-en.htm>

<sup>2</sup> Graduate School of Informatics,  
Kyoto University,  
Yoshida-honmachi, Sakyo-ku, 606-8501,  
Kyoto, Japan  
e-mail : inui@kyoto-u.ac.jp

**Abstract.** In previous research, we developed an integrated platform that combined visual scene interpretation with speech processing to provide input to a language learning model. The system was demonstrated to learn a rich set of sentence-meaning mappings that could allow it to construct the appropriate meanings for new sentences in a generalization task. While this demonstrated potential promise, it fell short in several aspects of providing a useful human-robot interaction system. The current research addresses three of these shortcomings, demonstrating the natural extensibility of the platform architecture. First, the system must be able not only to understand what it hears, but also to describe what it sees and to interact with the human user. This is a natural extension of the knowledge of sentence-to-meaning mappings that is now applied in the inverse scene-to-sentence sense. Secondly, we extend the system's ontology from physical events to include spatial relations. We will show that spatial relations are naturally accommodated in the predicate argument representations for events. Finally, because the robot community is international the robot should be able to speak multiple languages, and we thus demonstrate that the language model extends naturally to include both English and Japanese. Concrete results from a working interactive system are presented and future directions for adaptive human-robot interaction systems are outlined.

## 1 Introduction

As humanoid robots become increasingly capable of complex sensory and motor functions, the ability to interact with them in an ergonomic, real-time and adaptive manner becomes an increasingly pressing concern. In a previous study, we reported on progress in this direction, in terms of a system that could adaptively acquire a limited grammar based on training with human narrated video events. An overview of the system is presented in Figures 1. Figure 1A illustrates the physical setup in which the human operator performs physical events with toy blocks in the field of

view of a color CCD camera. Figure 1B illustrates a snapshot of the visual scene as observed by the image processing system. Figure 1C provides a schematic characterization of how the physical events are recognized by the image processing system.

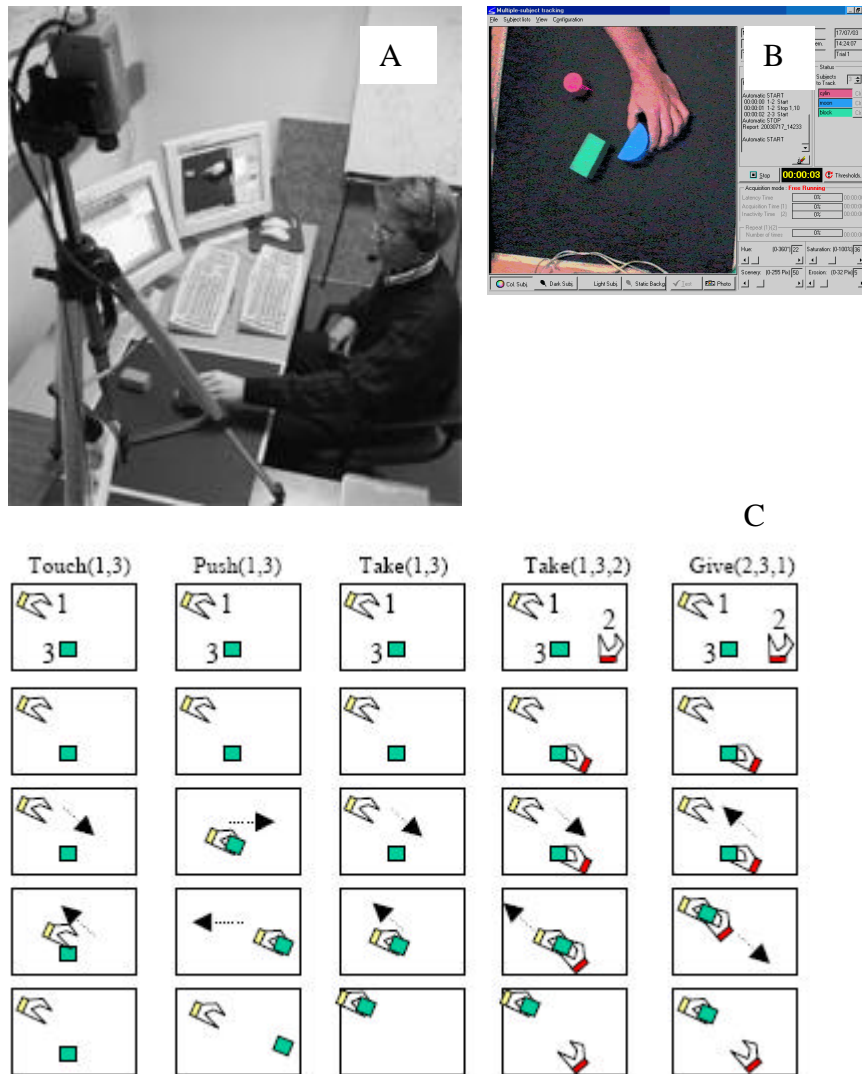


Figure 1. Overview of human-robot interaction platform. A. Human user interacting with the blocks, narrating events, and listening to system generated narrations. B. Snapshot of visual scene viewed by the CCD camera of the visual event processing system. C. Temporal contact sequence templates for recognition of touch, push, take and give events.

Using this platform, the human operator performs physical events and narrates his/her events. An image processing algorithm extracts the meaning of the events in terms of action(agent, object, recipient) descriptors. The event extraction algorithm detects physical contacts between objects, and then uses the temporal profile of contact sequences in order to categorize the events, based on the temporal schematic template illustrated in Figure 1C. While details can be found in Dominey (2003), the visual scene processing system is similar to related event extraction systems that rely on the characterization of complex physical events (e.g. give, take, stack) in terms of composition of physical primitives such as contact (e.g. Siskind 2001, Steels and Bailly 2003). Together with the event extraction system, a commercial speech to text system (IBM ViaVoice™) was used, such that each narrated event generated a well formed <sentence, meaning> pair.

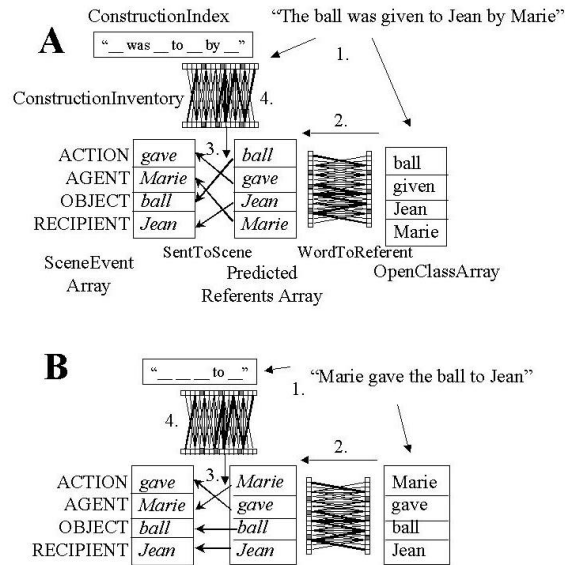


Figure 2. Grammatical construction architecture. Processing of active and passive sentence types in A, B, respectively. On input, Open class words populate the Open Class Array (OCA), and closed class words populate the Construction index. Visual Scene Analysis populates the Scene Event Array (SEA) with 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.

The <sentence, meaning> pairs were provided as training input to a learning model whose architecture is depicted in Figure2. The model integrates two powerful theories

of language acquisition to yield a robust learning capability. The essential problem the model is designed to address is that of mapping grammatical structure of sentences onto the semantic structure of their meanings. As illustrated in Figure 2 A and B, the problem of this mapping is not trivial, because a given language consists of a large ensemble of possible mappings. The first principle inherent in the model is that instead of representing <sentence, meaning> mappings in terms of a generative grammar, these mappings can be represented directly in a structured inventory of grammatical constructions that are nothing more than these mappings. Growing evidence both from studies of human language development (Tomasello 1999, 2003), and adult processing (Ferreira 2003, Sanford & Sturt 2002) indicate that a substantial component of language behavior can be accounted for in this manner. That is, that language production and comprehension is based on the re-use (including recombination) of existing templates, in a context in which the templates (i.e. grammatical constructions) can be learned by straight-forward mechanisms as illustrated in Figure 2. This does not exclude existence of truly generative mechanisms for construction and decoding new grammatical forms. However, for our purposes, in the domain of human-robot interaction, the ability to rapidly acquire relevant constructions in relatively restricted domains should prove quite useful. Indeed, as will be illustrated below, this capability will be of particular interest, as the meaning can include both descriptions, interrogations and commands.

If the language capability consists of a structured inventory of grammatical constructions, then the problem remains concerning how this inventory is managed. This is where the second great principle of developmental linguistics comes in: the cue competition hypothesis of Bates and MacWhinney (1982). They propose that across languages, there is a limited set of possible cues including word ordering regularities and the use of grammatical function words (e.g. to, by, from, that, was), that code the argument structure of sentences, that allows the determination of “who did what to whom.” Thus, as illustrated in Figure 2, the ensemble of closed class words together form "construction index" that serves as an index into an associative memory that stores the appropriate transformations. This memory store is referred to as the ConstructionInventory in Figure 2. In a series of experiments (Dominey 2003a, b) we have demonstrated that the system can thus learn an extensive set of grammatical constructions, including those in Table 1.

<b>Example Sentences and Meanings</b>	<b>Grammatical Constructions</b>
1. The block pushed the cylinder. Push(block, cylinder)	1. Agent verb object. (Active) Verb(agent, object)
2. The cylinder was pushed by the block. Push(block, cylinder)	2. Object was verbed by agent. (Passive) Verb(agent, object).
3. The block gave the cylinder to the moon. Give(block, cylinder, moon)	3. Agent verbed object to recipient. (Dative) Verb(agent, object, recipient)
4. The cylinder was given to the moon by the block.	4. Object was verbed to recipient by agent. (Dative passive)

Give(block, cylinder, moon)	Action1(agent1, object2, recipient3).
<b>Dual-Event Relative Constructions</b>	
6. The block that pushed the cylinder touched the moon. push(block, cylinder), Touch(block, moon)	6. Agent1 that verb1ed object2 verb2ed object3. Action1(agent1,object2), Action2 (agent1, object3)
7. The block was pushed by the moon that touched the cylinder. Touch(moon, cylinder), Push(moon, block)	7. Object3 was action2ed by agent1 that action1ed object2. Action1(agent1,object2), Action2 (agent1, object3)
17. The cat was given from the dog to the block that pushed the cylinder. Push(block, cylinder), Give(dog, cat, block)	17. Ag3 act2ed obj4 to recip1 that act1ed obj2 Action1(agent1,object2), Action2 (agent3,object4,recipient1)
18. The cylinder that was pushed by the block gave the cat to the dog. Push(block, cylinder), give(cylinder, cat, dog).	18. Obj4 was act2ed from ag3 to recip1 that act1ed obj2 Action1(agent1,object2), Action2 (agent3, object4,recipient1)
<b>Dual-Event Conjoined Constructions</b>	
27. The block pushed the cylinder and the moon. Push(block, cylinder), Push(block, moon)	27. Agent1 action1 object1 and object. Action1(agent1, object1), Action1(agent1, object2)
28. The block and the cylinder pushed the moon. Push(block, moon), Push(cylinder, moon)	28. Agent1 and agent3 action1ed object2. Action1(agent1, object2), Action1(agent3, object2)
29. The block pushed the cylinder and touched the moon. Push(block, cylinder), Touch(block, moon).	29. Agent1 action1ed object2 and action2 object3. Action1(agent1, object2), Action2(agent1, object3)
30. The moon and the block were given to the cylinder by the cat. Give(cat, moon, cylinder), Give(cat, block, cylinder).	30. Object2 and object3 were action1ed to recipient4 by agent1. Action1(agent1, object2, recipient4), Action1(agent1, object3, recipient4)

Table 1. Sample sentences with their meanings (left column) and the corresponding abstract grammatical constructions (right column).

## 2. Spoken Language Interaction

We thus demonstrated that the model could successfully learn 38 grammatical constructions, each of which allowed the system to generate the correct meaning for new sentences that had not been used in training. These initial learning results were quite promising, but of course the real test of utility is using this learned language capability in an interactive human-robot communication scenario. Technically there are several issues to be addressed, including (a) use of the learned grammatical constructions to generate sentences from visually perceived scenes, and to do so in a manner that is appropriate from a pragmatic discourse perspective; and (b) inserting this capability into an interactive environment coupled with speech synthesis and recognition.

### 2.1 Generating sentences from events.

Each grammatical construction in the construction inventory corresponds to a mapping from sentence to meaning. This information can thus be used to perform the inverse transformation from meaning to sentence. For the initial sentence generation studies we concentrated on the 5 grammatical constructions below. These correspond to constructions with two and three verb arguments in which each of the different arguments can take the focus position at the head of the sentence. On the left are presented example sentences, and on the right, the corresponding generic construction.

<u>Sentence</u>	<u>Construction</u>
1. The triangle pushed the moon.	<i>Agent event object.</i>
2. The moon was pushed by the triangle.	<i>Object was event by agent.</i>
3. The block gave the moon to the triangle.	<i>Agent event object to recipient.</i>
4. The moon was given to the triangle by the block.	<i>Object was event to recipient by agent.</i>
5. The triangle was given the moon by the block.	<i>Recipient was event object by agent.</i>

Table 2: Sentence and corresponding constructions for robot language generation.

This construction set provides sufficient linguistic flexibility, so that for example when the system is interrogated about the block, the moon or the triangle after describing the event *give(block, moon, triangle)*, the system can respond appropriately with sentences of type 3, 4 or 5, respectively. The important point is that each of these different constructions places the pragmatic focus on a different argument by placing it at the head of the sentence. Note that sentences 1-5 are specific sentences that exemplify the 5 constructions in question, and that these constructions each generalize to an open set of corresponding sentences. Thus, given an input meaning in the form *event(arg1, arg2, arg3)*, and an optional focus item (one of the three arguments), the system will deterministically choose the appropriate two or three argument construction, with the appropriate focus structure, in a pragmatically relevant manner.

## 2.2 Real-time Interactive Environment

The next task at hand is to integrate these pieces, including (a) scene processing for event recognition, (b) sentence generation from scene description and response to questions, (c) speech recognition for posing questions, and (d) speech synthesis for responding - into an interactive environment. The CSLU Speech Tools Rapid application Development (RAD) (<http://cslu.cse.ogi.edu/toolkit/index.html>) provides useful capability in this context. The system provides dialog management with a flexible and powerful graphical user interface with the global ability to link speech recognition and synthesis to the conditional execution of code on the same machine or on remote machines via ftp and socket protocols. This results in a hub architecture with RAD at the hub and the vision processing, language model, speech-to-text and voice synthesis at the periphery. Figure 3 illustrates the RAD flow of control configuration for our human-robot interaction demonstration. Here we briefly describe each state in the dialog management system and then provide a sample dialog.

1. *Start*: System initialization
2. *Welcome*: Says a verbal welcome to the user
3. *Object\_invitation*: Asks the user to move an object that the system will then name
4. *Get\_object\_description*: executes "an object from movement" algorithm on the "vision" file generated by the users movement. The visual system is triggered by any movement over a threshold velocity. It then records at a 5Hz sampling rate and continues until no motion occurs for at least 5 seconds, and generates an output file which triggers *Get\_object\_description*.
5. *Say\_object\_description*: executes speech synthesizer on the generated object description.
6. *Another\_object?*: invites user to show another object
7. *Want\_to\_go\_on?*: asks user if she wants to quit or go on
8. *Invitation*: invites user to perform an action with the blocks that the system will then describe
9. *Get\_description*: executes "describe event" on "vision" file. The output sentence is generated from the event description as described in 2.1 above.
10. *Say\_description*: executes synthesizer on the generated description
11. *Right?*: ask the user if the description was right, the user then responds that either the system was right (yes), it was wrong (no), or that there was system error (error). Each of these responses is processed separately and recorded in an SQL database for offline system evaluation.
12. *Any\_questions?*: invites the user to ask questions
13. *Block\_moon\_or\_cylinder?*: ask the user which of the objects she wants to know about. Based on the user response, the focus element is identified
14. *Get\_query\_response*: sends the focus element and the previous event description to the "respond-to-query" algorithm. This algorithm selects the correct grammatical construction to use depending on whether the event is of 2 or three arguments, and depending on the focus element (agent, object or recipient) as in 2.1.
15. *Say\_query\_reponse*: executes the synthesizer with the query response, and returns to *Any\_questions*.

16. *Again?:* offers the user to continue with another event (*make\_your\_move\_then*), or quit (*goodbye*).

17. *Make\_your\_move\_then:* re-invites the user to perform an event with the blocks.

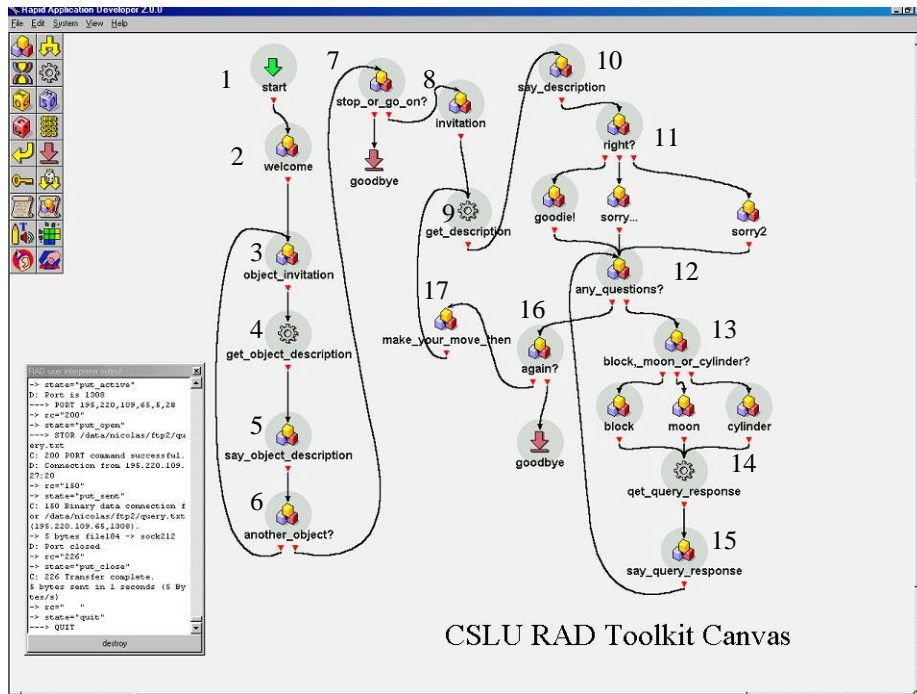


Figure 3. Rapid Application Development configuration for the Talking Robot Demonstration.

### 2.3 Example Dialog

Here we provide an example of typical dialog. Speech is indicated in *italics*.

Robot: *Welcome to the talking robot demonstration. Show me an object.*

Human: (Picks up and moves the block)

Robot: *That is the block. Another object?*

Human: *Yes* (moves the moon)

Robot: *That is the moon. Another object?*

Human: *No.*

Robot: *Shall we stop or go on?*

Human: *Go on.*

Robot: *When you are ready, make a movement with the objects so that I can see what you are doing. I will then describe what you did.*

Human: (picks up the block and uses it to push the cylinder to the moon).

Robot: *The block gave the cylinder to the moon. Was I correct?*



Human: *Yes.*  
 Robot: *Good for me. (stores results in database) Do you have any questions?*  
 Human: *Yes.*  
 Robot: *Would you like to know about the block, the moon or the cylinder?*  
 Human: *The moon.*  
 Robot: *The moon was gave the cylinder by the block. Do you have any questions?*  
 Human: *Yes.*  
 Robot: *Would you like to know about the block, the moon or the cylinder?*  
 Human: *The cylinder.*  
 Robot: *The cylinder was gave to the moon by the block.*  
 ....

Though the case is somewhat limited by the simplicity of the robot environment, this example provides a view of how language can provide a useful vector for human-robot interaction for describing and responding to questions about perceived events.

### 3. Extension of the ontology to Spatial Relations

Part of the claim to be supported by this research is that the current framework can be extended to satisfy the needs of different robot systems. 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. Central to this exercise is the idea that once a capability for mapping grammatical constructions to predicate-argument structures has been established for events, it should extend by analogy to any semantics that can be represented in predicate-argument format. 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  $\text{Left}(X,Y)$ ,  $\text{Above}(X,Y)$  etc. where  $X$  is the object that holds the spatial relation with the referent  $Y$ . That is,  $\text{Left}(X,Y)$  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  $\text{Relation}(X,Y)$  are available, where  $\text{Relation}$  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 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 objects, 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.

A

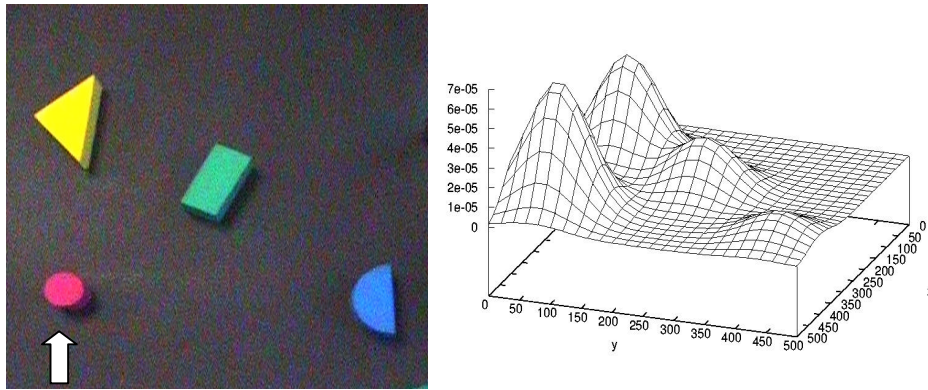


Figure 4. Spatial Attention for Relation Selection. The human user shows the robot a spatial relation and describes it. How does the robot know which of the multiple relations is the relevant one? A. The cylinder (lower left) has been moved into its current position, and now holds spatial relations with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

We collected data training data in which human subjects demonstrated and narrated spatial relations with the four objects. The spatial attention mechanism extracted for each case the most relevant spatial relation, and in well over 90% of the trails the attentional mechanism correctly selected the appropriate relation. The resulting <sentence, relation-meaning> pairs were used for training in the same procedure 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. This demon-

strates (1) the efficiency of the spatial attention mechanism, and (2) the generalizability of the predicate-argument approach.

#### 4. Extension to Japanese

While the construction framework can extend to new semantics on the meaning side, it would also be of significant value if it were able to extend to different languages. Indeed the construction framework is clearly cross-linguistic (Goldberg 1995). The current experiment will test the model with sentences in Japanese. Unlike English, Japanese allows extensive liberty in the ordering of words, with grammatical roles explicitly marked by postpositional function words -ga, -ni, -wo, -yotte. This word-order flexibility of Japanese with respect to English is illustrated here with the English active and passive di-transitive forms that each can be expressed in 4 different common manners in Japanese:

1. The block gave the circle to the triangle.
  - 1.1 Block-ga triangle-ni circle-wo watashita .
  - 1.2 Block-ga circle-wo triangle-ni watashita .
  - 1.3 Triangle-ni block-ga circle-wo watashita .
  - 1.4 Circle-wo block-ga triangle-ni watashita .
2. The circle was given to the triangle by the block.
  - 2.1 Circle-ga block-ni-yotte triangle-ni watasareta .
  - 2.2 Block-ni-yotte circle-ga triangle-ni watasareta .
  - 2.3 Block-ni-yotte triangle-ni circle-ga watasareta .
  - 2.4 Triangle-ni circle-ga block-ni-yotte watasareta .

In the “active” Japanese sentences, the postpositional function words -ga, -ni and -wo explicitly mark agent, recipient and, object whereas in the passive, these are marked respectively by -ni-yotte, -ga, and -ni. For both the active and passive forms, there are four different legal word-order permutations that preserve and rely on this marking. Japanese thus provides an interesting test of the model’s ability to accommodate such freedom in word order.

##### 4.1 Japanese Constructions

Each numbered sentence below is an example of a specific abstract grammatical construction type in Japanese whose meaning is provided in an event(argument) format *following* the sentence(s) corresponding to that meaning<sup>1</sup>. The corresponding

---

<sup>1</sup> hit = tatakuru, hit = tataita, be hit = tatakureru, was hit = tatakareta. give = ataeru, gave = watashita, be given = ataerareru, was given = watasareta. push = osu, pushed = tataita, be pushed = osareru, was pushed = osareta, believe = shinjiru, believed = shinjita. itself = jibun or jishin, it = sore.

English constructions are indicated in ()'s. Each construction can generalize to new sentences in which the open class elements are replaced.

- 1(1). block-ga circle-wo tataita.  
2(1). circle-wo block-ga tataita.  
*The block hit the circle.*  
*Hit(block, circle) active*
  - 3(2). Circle-ga block-ni tatakareta.  
4(2). Block-ni circle-ga tatakareta.  
*The circle was hit by the block.*  
*Hit(block, circle) passive*
  - 5(3). Block-ga triangle-ni circle-wo watashita .  
6(3). Block-ga circle-wo triangle-ni watashita .  
7(3). Triangle-ni block-ga circle-wo watashita .  
8(3). Circle-wo block-ga triangle-ni watashita .  
*The block gave the circle to the triangle.*  
*Gave(block, circle, triangle) active*
  - 9(4). Circle-ga block-ni-yotte triangle-ni watasareta .  
10(4). Block-ni-yotte circle-ga triangle-ni watasareta .  
11(4). Block-ni-yotte triangle-ni circle-ga watasareta .  
12(4). Triangle-ni circle-ga block-ni-yotte watasareta .  
*The circle was given to the triangle by the block.*  
*Gave(block, circle, triangle) passive*
  - 13(6). Circle-wo tataita block-ga triangle-wo oshita.  
*The block that hit the circle pushed the triangle.*  
*Hit(block, circle), Pushed(block, triangle)*
  - 14(7). Block-ga circle-wo oshita triangle-ni-yotte tatakareta.  
15(7). Circle-wo oshita triangle-ni-yotte block-ga tatakareta.  
*The block was hit by the triangle that pushed the circle.*  
*Pushed(triangle, circle), Hit(triangle, block)*
  - 16(8). Circle-wo tataita block-ga triangle-ni-yotte osareta.  
17(8). Triangle-ni-yotte circle-wo tataita block-ga osareta.  
*The block that hit the circle was pushed by the triangle.*  
*Hit(block, circle), Pushed(triangle, block)*
  - 18(9). Block-ga circle-wo oshita triangle-wo tataita.  
19(9). Circle-wo oshita triangle-wo block-ga tataita.  
*The block hit the triangle that pushed the circle.*  
*Pushed(triangle, circle), Hit(block, triangle)*
  - 20(10). Circle-ni-yotte tatakareta block-ga triangle-wo oshita.  
*The block that was hit by the circle pushed the triangle.*
  - 21(27). Block-ga circle-to triangle-wo tataita.
-

- 22(27). Circle-to triangle-wo block-ga tataita.  
*The block hit the circle and the triangle.*
- 23(28). Block-to triangle-ga circle-wo tataita.
- 24(28). Circle-wo block-to triangle-ga tataita.  
*The block and the triangle hit the circle.*
- 25(33). Block-ga sore-wo tataita triangle-wo oshita.  
*The block pushed the triangle that hit it.*
- 26(34). Block-ga sore-ga tataita triangle-wo oshita.  
*The block pushed the triangle that it hit.*

## 4.2 Learning Japanese Constructions

Employing the same method as described in the previous experiment, we thus expose the model to <sentence, meaning> pairs generated from the 26 Japanese constructions described below. We predicted that by processing the -ga, -ni, -yotte and –wo markers as closed class elements, the model would be able to discriminate and identify the distinct grammatical constructions and learn the corresponding mappings. Indeed, the model successfully discriminates between all of the construction types below; based on the ConstructionIndex unique to each construction type, and associates the correct SentenceToScene mapping with each of them. As for the English constructions, once learned, a given construction could generalize to new untrained sentences.

This demonstration with Japanese is an important validation that at least for this subset of constructions, the construction-based model is applicable both to fixed word order languages such as English, as well as free word order languages such as Japanese. This also provides further validation for the proposal of Bates and MacWhinney (1982) that thematic roles are indicated by a constellation of cues including grammatical markers and word order.

## 5. Status and future directions:

At the current writing, the spatial relations and the Japanese construction capabilities have been demonstrated but are not integrated into the interactive dialog system. These integrations are wholly feasible and will be realized in the future. This will provide a quite interesting bilingual capability that can provide a quite useful tool for human teams with English and Japanese speaking members collaborating together. Clearly, it will also have strong demonstration impact.

More generally, the near term future goal will be to export this system to a robot

platform that allows human-robot interaction not only about scene analysis but about action as well. This will provide the scenario in which language can be used to command and instruct the robot. Human based robot instruction has often relied on imitation, but clearly the use of verbal coaching and explaining will also provide a powerful information transfer mechanism. The current system has two important features that should make it of interest to the humanoid robot community. First, it is adaptable in the system will learn the language structures adapted to a given interaction context. Second, the system has a very flexible semantics in the form of predicate –argument relations. We have demonstrated that this is highly appropriate for event and spatial relation descriptions, but it will also be highly suitable for the syntax of robot commands, and should thus be of immediate practical value within the community.

## Acknowledgements

This work is funded by the French ACI Integrative and Computational Neuroscience, the HFSP, the European OMLL project. We gratefully acknowledge Nicolas Dermine who worked on the RAD integration for an earlier version of the system as part of an internship for the Lyon Ecole Central.

## 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 25<sup>th</sup> 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 language: I. Sensitivity to serial, temporal and abstract structure of language in the infant. *Lang. and Cognitive Processes*, 15(1) 87-127

- Dominey, P.F. (2003a) Learning Grammatical Constructions in a Miniature Language from Narrated Video Events, Proceedings of the 25<sup>th</sup> Annual Meeting of the Cognitive Science Society, Boston
- Dominey, P.F. (2003b) Learning Grammatical Constructions from Narrated Video Events for Human-Robot Interaction, Proc. IEEE Conf. On Humanoid Robotics, Karlsruhe.
- 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, (in press) *Brain and Language*
- 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 12<sup>th</sup> 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.
- Ferreira F (2003) The misinterpretation of noncanonical sentences, *cognitive Psychology*, 47, 164-203
- 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.
- Kotovskiy 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.
- Sanford AJ, Sturt P (2002) Depth of processing in language comprehension: not noticing the evidence. *Trends in Cognitive Science*, 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*, 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 probabilistic grammars by Bayesian model merging/ In *Grammatical Inference and Applications: Proc. 2<sup>nd</sup> 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
- Tomasello, M. (2003) *Constructing a language: A usage-based theory of language acquisition*. Harvard University Press, Cambridge.