

Language Learning For Human-Machine Interaction via Mapping of Grammatical Structure to Visual Scene Structure

Peter Ford Dominey

Abstract—Natural language interfaces will provide a privileged line of communication between humans and cognitive robots. In this context, the current research demonstrates that a 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 representations 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. 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. The benefits of the novel approach to scene analysis and language learning, and their interaction, are discussed.

Index Terms—perceptual scene analysis, language acquisition, model, neural network

I. INTRODUCTION

As aptly anticipated by Crangle and Suppes in 1993 [1], “robot technology is beginning to make its way off the factory floor and into our homes and places of work.” In an ever increasing manner, it will thus be necessary for humans to interact with robotic systems not only through formal programming and command languages, but also via natural language. While a significant degree of success has been realized through preprogrammed grammars [2], acquisition

of natural language interfaces through learning may offer the possibility of more robust and adaptive capabilities. More generally, autonomous robotic systems will challenge cognitive scientists and AI researchers to develop the systems that provide and enhance cognitive autonomy. Framed in another way, autonomous robotic platforms will permit the grounded testing of cognitive system models. From this perspective, the current research begins to address particular aspects of robot cognition based on vision and speech perception in the context of language acquisition and processing. This research proceeds in the tradition developed by Feldman & Lakoff and colleagues (reviewed in [3]), posed around the task in which “1. The system is given examples of pictures paired with true statements about those pictures in an arbitrary natural language. 2. The system is to learn the relevant portion of the language well enough so that given a novel sentence of that language it can determine whether or not the sentence is true of the accompanying picture.” Here, the task is posed in the context of dynamic scenes, with special attention to insights to be gained from knowledge of human cognitive development. In particular, these issues are addressed from the perspective of developmental aspects of human cognitive processing for language comprehension.

In the developmental trajectory of a human infant between 0-24 months of age the issues of perceptual scene analysis and natural language acquisition are addressed in a quite robust and effective manner. In particular, already at 6 months of age, infants are able to “parse” complex and dynamic visual scenes to identify causal events and their agents and goals, in a manner that is comparable to current machine vision methods [4,5]. 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 [6]. 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

This work was supported in part by the French Research Ministry under grants from the Cognitique Program on Action, and the Integrative and Computational Neuroscience initiative.

P. F. Dominey is with the Institut des Sciences Cognitives, CNRS UMR 5015, 67 Blvd. Pinel, 69675 Bron Cedex, France (phone: (33) 437911266 ; fax: (33) 437911210; e-mail: dominey@isc.cnrs.fr).

narration of scenes, and the internal representation of the analyzed scene.

II. GENERAL APPROACH

The generative grammar framework developed by Chomsky and his followers [7,8] has provided an extremely powerful analytical tool for the concise description of the grammatical structure of human languages. Indeed, central to the success of this program is the property that the “universal” description captures the richness of the grammatical structure of human languages. At the same time however, it must not be forgotten that an equally rich structure exists in the perceivable world, and that the communicative function of language is to transmit a linearized encoding of this structure from speaker to listener. This communication based approach to language has been less an issue for generative linguistic theorists than for cognitive scientists and engineers interested in building language based systems grounded in truly communicative contexts [9, 10].

The task of the listener, then, is to decode this linear sequence and reconstruct an internal representation equivalent to that which could be constructed from direct sensory perception. In this context, the current research is based on the construction or learning of mappings between grammatical structures and the structure of the perceptual scenes to which they correspond. Part of what the Chomsky program revealed is that sentences are not arbitrary collections of words, but rather, within each language, they are structured sequences that essentially unambiguously indicate “who did what to whom.” Thus, while the two sentences

1. The block pushed the triangle.
2. The triangle was pushed by the block.

are distinct linear sequences, they both correspond to the event *push(block, triangle)*. The current research is based on the principle that (a) Sentences are self-describing, and (b) The grammatical structure of sentences corresponds to the thematic structure of events in the visual scene. To the extent that these conditions are true, it should be possible to construct the mapping between grammatical structure in sentences, and thematic structure in visual scenes.

A. Sentences are Self-Describing

Part of the function of a language is to allow users to produce sentences that can be interpreted without ambiguity. In other words, the information necessary for thematic role assignment (i.e. determination of the bindings for *action(agent, object)*) is contained within the sentence itself. Thus, sentence (1) adheres to the “canonical” word ordering agent-action-object and can be unambiguously mapped onto

the situation in which “block” is the agent, “push” is the action and “triangle” is the object of the action. Similarly, sentence (2) is also self-describing, based on the modified word order and the grammatical function words “was,” (an auxiliary verb), and “by,” (a preposition). In these two examples, there is thus a fixed and completely well specified mapping between the grammatical structure of the sentence, and the corresponding thematic role assignment.

Concretely, a major assumption of the current approach is that sentences are a form of self-identifying data structure. That is, by relative combinations of word order conventions, explicit grammatical marking (by function words “by, to, from” etc. or grammatical morphemes), and their various combinations across languages, a given sentence in any language contains the information necessary to perform correct sentence-to-world mapping. Thus, an architecture that is sensitive to word order and grammatical marking shall provide the basis for a language-independent language acquisition system.

These limited observations can be extended such that all well formed sentences generated by a natural language grammar (such as that in section IV) will maintain this self-describing property, with a completely well specified mapping to the corresponding thematic role assignment. While this is clearly not the case for all natural language sentences, this restriction still allows a well posed and difficult problem of structure mapping between scenes and grammatical forms.

B. Grammatical Structure Maps to Thematic and Scene Structure

For sentences such as (1) or (2) above, the mapping from grammatical structure (subject-verb-object) onto thematic structure *action(agent, object)* is relatively straight-forward. Things become more complex for the relativised form in sentence (3).

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

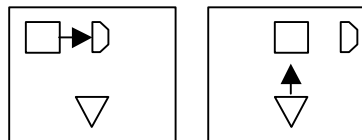


Figure 1. Visual scene events corresponding to the sentence “The block that pushed the moon was touched by the triangle.” The block first pushes the moon, and is then touched by the triangle.

Figure 1 illustrates successive frames of a visual sequence of events corresponding to sentence (3). These visual scene events can in turn be characterized from a thematic role perspective that corresponds to two successive events *push(block, moon)*, and *touch(triangle, block)*. Providing this level of structure in the scene representation allows the corresponding grammatical form of relative phrases (e.g. “.. the block *that pushed the moon*”), addressing a major

limitation of sentence and scene processing systems identified by Feldman [3]. That is, the grammatical complexity of sentences will be limited by the structural complexity of represented scenes. Here, the “dual scene” representation allows grammatical complexity including the use of relative phrases.

More generally, all well formed sentences generated by the grammar in section IV will maintain this self-describing property, i.e. they will have a completely well specified mapping to the corresponding thematic role assignment. The intermediate step that remains is the extraction of thematic roles in an event description based on analysis of actual dynamic visual scenes.

C. Extracting Events from Dynamic Visual Scenes

The visual scenes are made up of actions corresponding to *touch*, *push*, *take* and *give* that occur with colored toy blocks on an experimental workspace. The objects are manipulated by the experimenter who at the same time narrates the ongoing events. The visual images are captured by a standard CCD camera, and then processed by a commercial video image analysis system that yields a time-ordered list of the physical contacts between objects, and associated parameters including the relative velocities of the objects involved in a contact, and the duration of the contact. Based on this perceptual primitive information, a higher level representation is constructed in terms of specific events (touch, push, take, give) and the associated agent, object and recipient for each event. In parallel, the speech narrative of the ongoing events is processed by commercial voice recognition software, to generate a text file of the narrative. Together, the set of perceptual events, and the corresponding set of event narratives are provided as input to an associative structure mapping algorithm that learns the mapping between words and their referents in the scene, and between grammatical structures and their event-level interpretations in the scene.

III. VISUAL SCENES AND ANALYSIS

From the perspective of human cognitive development, it is well known that among the most early acquired perceptual capacities is that of detecting physical contact between objects [11]. Based on this finding, contact will be treated as a perceptual primitive upon which more complex event representations will be built in visual scene analysis.

The visual environment consists of three objects on a black matte surface. The objects are a red cylinder, a green block and a blue semicircle or “moon.” Visual scenes are made up of events that occur between these objects, physically generated by the experimenter. The simplest events, *touch*, *push*, and *take*, involve the causal agent, and the object. The event *give* involves a third role corresponding to the recipient, and take can also involve a third argument in the case that the

object is being taken from another object.

Within this context, the objective of the visual scene analysis is, for a given video sequence, to generate the corresponding event description in the format $event_i(agent, object, recipient)$, where event corresponds to touch, push, take or give, and $i = 1$ for simple events (e.g. corresponding to sentences (1) and (2)), and $i = (1, 2)$ for compound events (e.g. corresponding to sentence (3)).

A Sony CCD camera is located 85 cm above the surface with a field of view of 1 meter in diameter. The video image is processed by a color-based analysis system (Smart – Panlab, Barcelona Spain) that generates a time ordered sequence of the contacts that occur between objects in the field of view of the camera. Based on this contact sequence, the higher level event description is extracted.

A. Single Event Labeling

Scene events are defined in terms of contacts between elements. 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 [4]. Based on these parameters of contact, scene events are recognized as follows:

1) *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.

2) *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.

3) *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.

4) *Take(agent, object, source)*

In this event, the agent takes the object from the source. This is a compound event in that it is made up of multiple contacts. For the first contact between the agent and the object (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. For the optional second contact between the agent and the 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.

5) *Give(agent, object, recipient)*

In this event, the agent first takes the object, and then gives the object to the recipient. This is a compound event, made up of multiple contacts. 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.

These event labeling templates form the basis for a template matching algorithm that labels events based on the contact list. Initial studies identified events based on the agency (determined as the element with the maximum relative velocity), and the contact durations. Displacement parameters can provide a more robust characterization for event discrimination.

B. Complex “Hierarchical” Events

The events described above are simple in the sense that there have no recursive or hierarchical structure. This imposes serious limitations on the syntactic complexity of the corresponding sentences [4]. Figure 1 illustrates a complex event that illustrates this issue. Such a compound event will be recognized and represented as a pair of temporally aligned 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.

IV. LANGUAGE AND LEXICAL ANALYSIS

At the same time that the experimenter generates these visual scene events by manipulating objects in the field of view of the camera, he/she simultaneously narrates the ongoing events. This narrative is processed by a commercial voice recognition system (IBM ViaVoice™). An off-line analysis of approximately 300 sentences generated under these conditions revealed that the subset of English that was being employed can be described by the following relatively simple context free grammar:

- | | |
|----------------------------|-----------------------------|
| 1. S → NP + VP | [sentence] |
| 2. NP → Det + N | [noun phrase] |
| 3. NP → NP + Rel + VP | [relative noun phrase] |
| 4. VP → Va + NP | [active verb phrase 1 arg] |
| 5. VP → Va + NP + PP | [active verb phrase 2 arg] |
| 6. VP → Aux + Vp + PP | [passive verb phrase 1 arg] |
| 7. VP → Aux + Vp + PP + PP | [passive verb phrase 2 arg] |

- | | |
|--|------------------------|
| 8. PP → Prep + NP | [prepositional phrase] |
| 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 | |
| 15. Rel → that | |

This grammar allows generation of sentences such as: “The cylinder that was touched by the moon pushed the triangle.” While this grammar is clearly a reduced subset of English, it is rich enough to provide a challenge to any language learning system, particularly through its inclusion of passives and relative noun phrases.

Because of the underlying structure, sentences generated by this grammar will have certain interesting properties as mentioned above. In particular, thematic roles are uniquely assigned by grammatical structure. In many languages, including English, this grammatical structure is instantiated in terms of regularities on word order, and by a special category of words, referred to grammatical function words, including determiners, prepositions, auxiliary verbs, as described above in IIA and B.

This lexical categorization of function words vs content words (i.e. those words that carry a semantic content, including nouns, verbs, adjectives and adverbs) is the foundation of language comprehension. Indeed, it has been behaviorally demonstrated that newborn infants are sensitive to the acoustic and statistical distribution properties that distinguish these two categories [12], and in adults, these two word categories are processed by dissociable neurophysiological systems [13]. Similarly, it has been demonstrated that artificial neural networks can also learn to make this function/content distinction [12, 14]. Based on the fundamental importance of this distinction, and the demonstration that this categorization can be learned, the speech input that is provided to the learning model will undergo one “pre-processing” step that corresponds to an explicit marking of the function vs. content status of each word.

V. STRUCTURE MAPPING FOR LANGUAGE LEARNING

As mentioned in the introduction, the learning task is quite similar to the “picture – narration” learning task posed by Feldman et al [4] in which a language is to be learned based on scene – narration pairs. The current version slightly extends this to dynamic visual scenes rather than pictures. Feldman et al [4] observed that as the task is posed, there is no explicit requirement to induce the underlying grammar – the agent could learn the mappings between surface forms and semantics of the scene representations without any notion of syntax. We take a related but significantly different

approach in considering that the learned mapping *actually corresponds to* the syntactic structure of the language.

The mapping of sentence structure onto scene event structure can be considered at two distinct levels. At one level, words are associated with individual components of event descriptions. At the second level, grammatical structure is associated with functional roles within scene events. The first level has been addressed elsewhere [15, 16], 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.

A. Initial Processing of Vision and Speech

The structure mapping architecture is illustrated in Figure 2. As described above, visual scene processing for a given scene generates a scene description. This description is encoded in the Scene Event Array (SEA) that consists of two sub-arrays so as to accommodate complex events (see IIIB). Each sub-array contains fields corresponding to *action*, *agent*, *object*, *recipient/source*. Each field is a 25-element vector with a single bit-on encoding. The SEA thus allows representation of the five simple event types, as well as their combinations in hierarchical events.

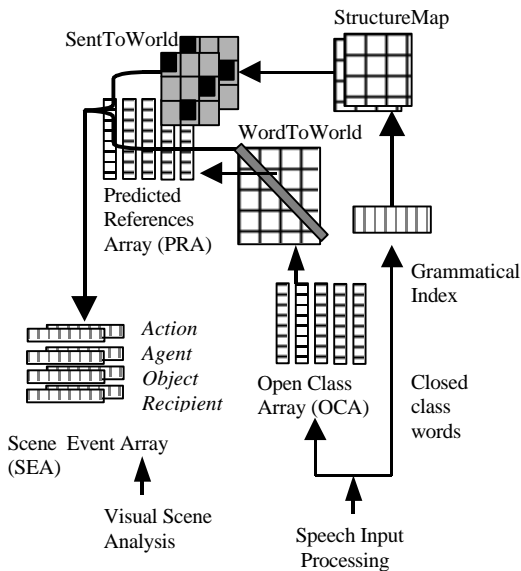


Figure 2. Structure-Mapping Architecture for language learning. Open class words in OCA are translated to scene elements in the PRA via the WordToWorld mapping. The PRA elements are then mapped onto their respective roles in the scene SEA by the SentenceToWorld mapping, specific to each sentence type, retrieved from StructureMap, via the GrammaticalIndex that encodes the closed class function words that characterize each sentence type.

Similarly, the processing of voice narrative yields an ASCII text representation. As described above (in IV) function and content words will be represented in distinct manners. Content words are coded with single bit-on in the range 1-16, and function words in 17-25. The content (or open class) words will be encoded in the Open Class Array (OCA) that

contains 6 fields, each a 25-element vector with single bit-on encoding. The function (or closed class) words are encoded in a linear array called the Grammatical Index described below.

B. Word Meaning

In the initial learning phases, the association between a word and its corresponding scene item is learned by a simple associative memory, and is stored in the WordToWorld matrix (Eqn 1). In Eqn 1, the index $k = 1$ to 6, corresponding to the maximum number of words in the open class array (OCA). Index $m = 1$ to 6, corresponding to the maximum number of elements in the scene event array (SEA). Indices i and $j = 1$ to 25, corresponding to the word and scene item vector sizes, respectively. LRSem is a learning rate parameter. In this initial configuration the term α is zero, and this learning simply associates every word with every element in the current scene. This exploits a form of cross situational learning, in which the correct word-scene item associations will emerge as that which remains common across multiple sentence-scene situations [16]. In this manner the system can extract the cross-situational regularity that a given word will have a higher coincidence with the scene object to which it refers than with other objects. This allows initial word learning to occur, which contributes to learning the mapping between sentence and scene structure (Eqn. 4, 5 & 6 below). Once this learning has occurred, knowledge of the syntactic structure, encoded in SentenceToWorld can be used to identify the appropriate referent (in the SEA) for a given word (in the OCA), corresponding to a non-zero value of α in Eqn. 1. This corresponds to a form of “syntactic bootstrapping” in word learning. Thus, for the new word “gugle”, syntactic knowledge of the sentence “John pushed the gugle” can be used to assign “gugle” to the object of push.

$$\text{WordToWorld}(i,j) = \text{WordToWorld}(i,j) + \text{OCA}(k,i) * \text{SEA}(m,j) * \text{LRSem} * \alpha \text{SentenceToWorld}(m,k) \quad (1)$$

C. Mapping Sentence to Scene

As stated above (II), for every sentence in our language, there exists a corresponding scene such that there is a perfect mapping from open class elements in the sentence onto event elements in the scene. In terms of the architecture in Figure 2, this can be restated in the following successive steps. First, words in the Open Class Array are decoded into their corresponding scene referents (via the WordToWorld mapping) to yield the Predicted Referents Array (Eqn 2) that contains the translated words while preserving their original order from the OCA. Index $k = 1$ to 6, corresponding to the maximum number of scene items in the predicted referents array (PRA). Indices i and $j = 1$ to 25, corresponding to the word and scene item vector sizes, respectively.

$$\text{PRA}(k,j) = \sum_{i=1}^n \text{OCA}(k,i) * \text{Word-to-World}(i,j) \quad (2)$$

Next, each grammatical form generated by the language will correspond to a specific mapping between the PRA and the SEA. Distinct instances of these mappings are encoded in the SentenceToWorld array for the different grammatical forms. The problem will be to retrieve for each grammatical form, the appropriate corresponding SentenceToWorld mapping. To solve this problem, we ensure that each grammatical form will have a unique corresponding Grammatical Index. Thus, the appropriate SentenceToWorld mapping for each grammatical form can be indexed by its corresponding Grammatical Index.

The Grammatical Index (Eqn.3) encodes the function words of a sentence, preserving their order of arrival and their relative position in the sentence. Since each grammatical form has a unique configuration of function words, with respect to their identity, order and relative position (IIA), the Grammatical Index will thus uniquely identify each distinct grammatical form. The Grammatical Index is a 25 element vector. Each function word is encoded as a single bit in a 25 element FunctionWord vector. When a function word is encountered during sentence processing, the current contents of Grammatical Index are shifted by $n + m$ bits in a circular shift (indicated by f_{cs}) where n corresponds to the bit that is on in the FunctionWord, and m corresponds to the number of open class words that have been encountered since the previous function word (or the beginning of the sentence). Finally, a vector addition is performed on this result and the FunctionWord vector.

$$\begin{aligned} \text{Grammatical Index} &= f_{cs}(\text{Grammatical Index}, (n + m)) \\ &+ \text{FunctionWord} \end{aligned} \quad (3)$$

GrammaticalIndex thus encodes the function words, their relative order and their relative position with respect to the content words. The link between the Grammatical Index and the corresponding SentenceToWorld mapping is established as follows. As each new sentence is processed, we first reconstruct the specific SentenceToWorld mapping for that sentence (Eqn 4). The resulting, SentenceToWorldCurrent 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 SentenceToWorldCurrent will depend on the quality of acquired word meanings as reflected in the PRA. Thus, syntactic learning requires a minimum baseline of semantic knowledge. Index $m = 1$ to 6, corresponding to the maximum number of elements in the scene event array (SEA). Index $k = 1$ to 6, corresponding to the maximum number of words in the predicted references array (PRA). Index $i = 1$ to 25, corresponding to the word and scene item

vector sizes.

Given the SentenceToWorldCurrent mapping for the current sentence, we can now associate it with the corresponding function word configuration for that sentence, expressed in the Grammatical Index (Eqn 5). Note that we have linearized SentenceToWorld-Current from 2 to 1 dimensions to make the matrix multiplication more transparent. Thus index j varies from 1 to 36 corresponding to the 6x6 dimensions of SentenceToWorldCurrent.

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

$$\begin{aligned} \text{StructureMap}(i,j) &= (\text{StructureMap}(i,j) \\ &+ \text{Grammatical Index}(i) \\ &* \text{Sentence-to-World-Current}(j) \\ &* \text{LRSyn}) / \text{Sum}(\text{StructureMap}) \end{aligned} \quad (5)$$

Finally, once this learning has occurred, for new sentences we can now extract the SentenceToWorld mapping from the learned StructureMap by using the Grammatical Index as an index into this associative memory, illustrated in Eqn. 6. Again to simplify the matrix multiplication, Sentence-to-World has been linearized to one dimension, based on the original 6x6 matrix. Thus, index $i = 1$ to 36, and index $j = 1$ to 25 corresponding to the dimension of the Grammatical Index.

$$\begin{aligned} \text{SentenceToWorld}(i) &= \\ &\sum_{i=1}^n \text{StructureMap}(i,j) * \text{GrammaticalIndex}(j) \end{aligned} \quad (6)$$

Note that the learning in (5) encodes the association between GrammaticalIndex and the desired Sentence-to-World mapping in StructureMap. We can also assure that false associations are not encoded, by using the error between the “correct” SentenceToWorldCurrent and estimated SentenceToWorld from (6). This error can then be used to weaken the association between GrammaticalIndex and appropriate SentenceToWorld elements.

To accommodate the dual scenes for complex events (III), 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.

We evaluate performance by using the WordToWorld and SentenceToWorld 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 Sentence-to-World transformation as described in Eqn. 7.

$$PSA(m,i) = PRA(k,i) * SentenceToWorld(m,k) \quad (7)$$

In Eqn 7, index $i = 1$ to 25 corresponding to the size of the scene and word vectors. Indices m and $k = 1$ to 6, corresponding to the dimension of the predicted scene array, and the predicted references array, respectively. 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.

VI. EXPERIMENTAL RESULTS

In the following experiments the human operator manipulated colored toy blocks in the CCD visual field, and simultaneously narrated his actions. Speech and vision data were acquired and then processed off-line. The two data streams were temporally aligned so that corresponding scene and narration were paired for each event. Each experiment thus yielded a data set of matched sentence – scene pairs that were provided as input to the structure mapping model.

A. Initial Learning of Active Forms for Simple Events

The first experiment examined learning with the five event types, and corresponding narrations only in the active voice, corresponding to the grammatical forms 1 and 2.

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. The model was trained for 32 passes through the 17 scene/sentence pairs for a total of 544 scene/sentence pairs (Fig 3, Exp A). During the first 200 trials (scene/sentence pairs), value α in Eqn. 1 was, 0, and thereafter it was 1. This was necessary in order to avoid the random effect of syntactic knowledge on semantic learning in the initial learning stages. Evaluation of the performance of the model after this training indicated that for all 17 sentences, there was error-free performance. A clear test of language learning is the ability to generalize to new sentences that have not previously been tested. Generalization in this form also yielded error free performance.

B. Passive forms

The second experiment examined learning with the five event types and the introduction of passive grammatical forms, thus employing grammatical forms 1-4.

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 that employed the different event types, with two- and three-arguments, and active and passive grammatical forms for the narration. Word meanings were used from Experiment A, so only the structural mapping from grammatical to scene structure was learned. As indicated in Figure 3 (Exp B), within 3 training passes through the 17 sentences, error free performance was achieved. Note that only the WordToWorld mappings were retained from Experiment A. Thus, the 4 grammatical forms were learned from the initial naive state. In the generalization test, the learned values were fixed, and the model demonstrated error-free performance on new sentences for all four grammatical forms that had not been used during the training.

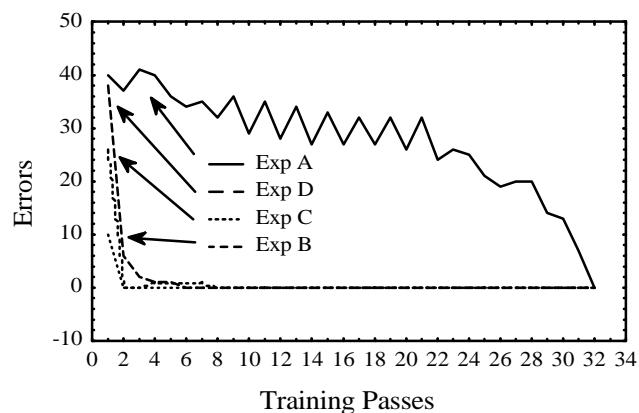


Figure 3. Evolution of interpretation errors during learning. Once the lexicon has been developed in Exp A, interpretation errors are significantly reduced despite the increase of syntactic complexity in Exp C and D. In these cases, error free performance is realized in 32, 2, 8 and 6 passes through the training corpus, respectively.

In this experiment, only 2 grammatical forms were learned, and the lexical mapping of words to their scene referents was learned. As stated in part VC, word meaning provides the basis for extracting more complex syntactic structure. Thus, these word meanings are fixed and used for the subsequent experiments.

C. Relative forms for Complex Events

The complexity of the scenes and corresponding grammatical forms in the previous experiments were quite simple. Here we consider complex scenes as illustrated in Figure 1. 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 8 presentations of the 11 scene/sentence training sentences, the model performed without error for these 6 grammatical forms. In the generalization test, the learned values were fixed, and the model demonstrated error-free performance on new sentences for all six grammatical forms that had not been used during the training.

D. Combined Test

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 B and C, were employed that exercised the ensemble of grammatical forms 1 – 10. 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.

VII. DISCUSSION

The stated objective of this research was to demonstrate that an autonomous 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. The results presented in section VI indicate that this objective has been achieved, and here, certain strengths and weaknesses will be addressed.

A. Visual Scene Processing

One of the achievements of this research has been the development of a system for event extraction based on two fundamental system design decisions. The first decision was to modularize scene processing, so that object recognition and tracking are separated from event recognition. This allows the use of a robust commercial recognition and tracking system (SmartTM), that provides position and relative proximity (contact) data for the objects. The second decision was to then perform event recognition based on studies of human infants' event perception. Particularly, contact and the associated parameters of duration and displacement provide highly robust cues for allowing the categorization and agent identification of simple events [4, 5, 11].

B. Word Meaning

In the current research we admittedly have taken a simplified approach to learning the meanings of individual words. This is not a particular problem, however, as we know that this problem can be resolved by more advanced methods as demonstrated by Roy and Pentland [15], and by Siskind [16], while our current effort was more specifically concentrated on the structural mapping between visual scene descriptions and the natural language grammar of sentences describing those scenes.

C. Grammatical to Scene Structure Mapping

The problem of language understanding has been reformulated, based on [4], as the problem of learning a generalized mapping from sentences onto scenes through training on a limited set of sentences that characterize (a subset of) the language. The results of experiments A-D indicate that once a baseline of words have been acquired, the use of this knowledge for acquisition of new grammatical forms is quite robust and rapid. There are at least two reasons for this learning ability that are noteworthy.

Firstly, the current approach does not explicitly try to learn the rules of the underlying grammar. Rather, it learns for each grammatical structure (sentence type) the mapping between the open class words in that sentence and their role in the scene event array. To the extent that sentences are explicit about "who did what to whom", the model will be able to exploit this explicit coding and use it to retrieve the appropriate mapping.

This leads to the second point of interest concerning how this mapping can be retrieved. Indeed, the problem can be further refined to the problem of extracting from each sentence-type a unique "grammatical index" that can be used to retrieve the appropriate mapping. We hypothesized that such an index could be created based on the processing of the grammatical function elements in the sentence. In English, these functional elements are typically instantiated as function words, though the general case should also handle functional morphemes attached to the open class words. This mechanism based on the grammatical index should thus be sufficient for the general acquisition problem. Interestingly however, as in the scene processing, this mapping has been modularized such that the nature of the grammatical index formation may evolve, independently of the mapping process, and vice-versa.

D. Structure in the Scene

If grammatical structure is to map to scene structure, then the grammar must be of sufficient complexity to match that of the scene. In the current study, we demonstrated that by introducing a "dual" or "complex" scene representation, we can accommodate grammatical forms with recursive structure in the form of relative phrases (e.g. The triangle that pushed the ball took the block). More generally, the complexity of

the represented scene must be matched by that of the target language. Thus, the problem of grammatical complexity is in a sense turned around, and shown to derive directly from the complexity of the scene.

E. Relevance

In 1968-1970 Terry Winograd developed SHRDLU, a Lisp-based A.I. system that understood natural language and carried on (typed) dialogs with human users about a small world of blocks on a table surface [17]. The system was extremely powerful in its restricted domain, and represented an impressive combination of natural language processing and problem solving. The system was not adaptive, however, and did not appear to scale well to larger problem domains, particularly those in which new linguistic forms would be used, and in which dynamic events, rather than static configurations were to be manipulated. This adaptive aspect of the problem was subsequently posed as a challenge by Feldman and colleagues [3].

Within this context, the current research is of relevance for two distinct but complementary reasons. Firstly, it proposes a new and effective method for such autonomous systems to acquire knowledge of a natural language grammar by exposure to visual scenes, and speech that narrates those scenes. It achieves this in an integrated, end-to-end system in which real visual events and speech are analyzed and then processed together in a novel structure-mapping architecture. The system is adaptive, and thus learns the crucial relations between grammatical structure and event structure, thus escaping from the scaling limitation of using a preprogrammed grammar.

The second interest of this work is that it exploits extensive knowledge of the human developmental trajectory, and adult processing of language and scenes from a systems perspective. The language processing model is an extension of a family of language processing models that were developed based on the neurophysiology of human language and cognitive sequence processing [18, 19]. Likewise, the scene analysis and event extraction methods are based on data from human developmental studies [4,5,11]. In particular, the system exploits the use of perceptual primitives that are highly salient (i.e. contact) in order to characterize complex events such as give and take. This is of interest because in effect, it demonstrates that this cybernetic approach can be used to more realistically simulate the developmental trajectory of human cognition [18]. Indeed, such an approach shall contribute both to the development of more robust autonomous systems, and to the understanding of human cognition.

REFERENCES

- [1] Crangle C. & Suppes P. (1994) Language and Learning for Robots, CSLI lecture notes: no. 41, Stanford.
- [2] Dowding J, Gawron JM, Appelt D, Bear J, Chernay L, Moore R, Moran D (1993) Gemini: A natural language system for spoken-language understanding. Proc. 31st Ann Mtg Assoc Comp. Linguistics, 1993, p. 54-61.
- [3] J. Feldman, G. Lakoff, D. Bailey, S. Narayanan, T. Regier, A. Stolcke (1996). LO: The First Five Years. Artificial Intelligence Review, v10 103-129.
- [4] Leslie AM, Keeble S (1987) Do six-month-olds perceive causality? Cognition 25, 265-288.
- [5] Woodward AL (1998) Infants selectively encode the goal object of an actor's reach. Cognition 69 1-34.
- [6] Hirsh-Pasek K, Golinkoff RM (1996) The origins of grammar: evidence from early language comprehension. MIT Press, Boston.
- [7] Chomsky N. (1995) The Minimalist Program. MIT
- [8] Craine S, Lillo-Martin D, An introduction to linguistic theory and language acquisition, 1999, Blackwell
- [9] Langacker, R. (1991). Foundations of Cognitive Grammar. Practical Applications, Volume 2. Stanford University Press, Stanford.
- [10] Steels, L. (2001) Language Games for Autonomous Robots. IEEE Intelligent Systems, vol. 16, nr. 5, pp. 16-22, New York: IEEE Press.
- [11] Kotovsky L, Baillargeon R, The development of calibration-based reasoning about collision events in young infants. 1998, Cognition, 67, 311-351
- [12] 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.
- [13] 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
- [14] Morgan JL, Shi R, Allopenna P (1996) Perceptual bases of rudimentary grammatical categories: Toward a broader conceptualization of bootstrapping, pp 263-286, in Morgan JL, Demuth K (1996) Signal to syntax: bootstrapping from speech to grammar in early acquisition. Lawrence Erlbaum, Mahwah NJ, USA.
- [15] Deb Roy and Alex Pentland. (2002). Learning Words from Sights and Sounds: A Computational Model. Cognitive Science, 26(1), 113-146.
- [16] Siskind JM (1996) A computational study of cross-situational techniques for learning word-to-meaning mappings, Cognition (61) 39-91.
- [17] Winograd T (1972) Understanding Natural Language, Academic Press
- [18] Dominey PF (2000) Conceptual Grounding in Simulation Studies of Language Acquisition, *Evolution of Communication*, 4(1), 57-85.
- [19] 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*

Peter Ford Dominey (M'88) completed the BA at Cornell University, Ithaca NY in 1984 in the College Scholar Program, in the fields of cognitive psychology and artificial intelligence. In 1989 and 1993 respectively he obtained the M.Sc. and Ph.D. in computer science from the University of Southern California, Los Angeles CA, developing neural network models of sensorimotor sequence learning.

From 1984 to 1986 he was a Software Engineer at The Data General Corporation in Westboro MA, and from 1986 to 1993 he was a Systems Engineer at the Jet Propulsion Laboratory in Pasadena CA. From 1993 to 1997 he was a post-doctoral fellow at the Vision and Motor Control laboratory of INSERM in Lyon France, and in 1997 he became a tenured researcher in the Centre National de la Recherche Scientifique (CNRS). He currently leads a research group on Sequential Cognition and Language at the Institut des Sciences Cognitives in Lyon France. His research interests include understanding and simulating the neurophysiology of cognitive sequence processing and language, and the application of this knowledge to humanoid robot cognition and language processing.