Learning Grammatical Constructions from Narrated Video Events for Human-Robot Interaction

Peter Ford Dominey
Institut des Sciences Cognitives, CNRS
67 Blvd. Pinel,
69675 Bron Cedex, France
dominey@ isc.cnrs.fr

Abstract. The ability to provide natural language interaction with humanoids will likely require adaptive systems that learn the mappings from speech to meaning in a context dependant manner. The objective of this research is to develop a system for miniature language learning based on a minimum of pre-wired language-specific functionality, that is compatible with observations of perceptual and language capabilities in human development. In the proposed system, meaning is extracted from video images based on detection of physical contact and its parameters. Mapping of sentence form to meaning is performed by learning grammatical constructions that are retrieved from a construction inventory based on the constellation of closed class items uniquely identifying the target sentence structure. The resulting system displays robust acquisition behavior that reproduces certain observations from developmental studies, with very modest "innate" language specificity.

1 Introduction

As aptly anticipated by Crangle and Suppes [5] in 1994, "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, 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. Feldman et al. [10, 11] posed the problem of "miniature" language acquisition based on <sentence, image> pairs as a "touchstone" for cognitive science. In this task, an artificial system is confronted with a reduced version of the problem of language acquisition faced by the child, that involves both the extraction of meaning from the image, and the mapping of the paired sentence onto this meaning.

1.1 Extraction of Meaning

In this developmental context, Mandler [17] suggested that the infant begins to construct meaning from the scene based on the extraction of perceptual primitives. From simple representations such as contact, support, attachment [27] the infant could construct progressively more elaborate representations of visuospatial meaning. Thus, the physical event "collision" is a form of the perceptual primitive “contact”. Kotovsky & Baillargeon [15] observed that at 6 months, infants demonstrate sensitivity to the parameters of objects involved in a collision, and the resulting effect on the collision, suggesting indeed that infants can represent contact as an event predicate with agent and patient arguments.

Siskind [25] has demonstrated that force dynamic primitives of contact, support, attachment can be extracted from video event sequences and used to recognize events including pick-up, put-down, and stack based on their characterization in an event logic. The use of these intermediate representations renders the system robust to variability in motion and view parameters. Most importantly, Siskind demonstrated that the lexical semantics for a number of verbs could be established by automatic image processing.

1.2 Sentence to meaning mapping:

Once meaning is extracted from the scene, the significant problem of mapping sentences to meanings remains. The nativist perspective on this problem holds that the <sentence, meaning> data to which the child is exposed is highly indeterminate, and underspecifies the mapping to be learned. This “poverty of the stimulus” is a central argument for the existence of a genetically specified universal grammar, such that language acquisition consists of configuring the UG for the appropriate target language [3]. In this framework, once a given parameter is set, its use should apply to new constructions in a generalized, generative manner.

An alternative functionalist perspective holds that learning plays a much more central role in language acquisition. The infant develops an inventory of grammatical constructions as mappings from form to meaning [12]. These constructions are initially rather fixed and specific, and later become generalized into a more abstract compositional form employed by the adult [28]. In this context, construction of the relation between perceptual and cognitive representations and grammatical form plays a central role in learning language (e.g. [10, 11, 16, 17, 27, 28]).

These issues of learnability and innateness have provided a rich motivation for simulation studies that have taken a number of different forms. Elman [9] demonstrated that recurrent networks are sensitive to predictable structure in grammatical sequences. Subsequent studies of grammar induction demonstrate how syntactic structure can be recovered from sentences (e.g. [26]). From the “grounding of language in meaning” perspective (e.g. [10, 11, 16, 17, 27]) Chang & Maia [4] exploited the relations between action representation and simple verb frames in a construction grammar approach. In effort to consider more complex grammatical forms, Miikkulainen [18] demonstrated a system that learned the mapping between relative phrase constructions and multiple event representations, based on the use of a stack for maintaining state information during the processing of the next embedded clause in a recursive manner.

In a more generalized approach, Dominey [7] exploited the regularity that sentence to meaning mapping is encoded in all languages by word order and grammatical marking (bound or free) [1]. That model was based on the functional neurophysiology of cognitive sequence and language processing and an associated neural network model that has been demonstrated to simulate interesting aspects of infant [6] and adult language processing [8].

1.3 Objectives

The goals of the current study are fourfold: First to test the hypothesis that meaning can be extracted from visual scenes based on the detection of contact and its parameters in an approach similar to but significantly simplified from Siskind [24]; Second to determine whether the model of Dominey [7] can be extended to handle embedded relative clauses; Third to demonstrate that these two systems can be combined to perform miniature language acquisition; and finally to demonstrate that the combined system can provide insight into the developmental progression in human language acquisition without the necessity of a pre-wired parameterized grammar system [3].

1.4 The Training Data

The human experimenter enacts and simultaneously narrates visual scenes made up of events that occur between a red cylinder, a green block and a blue semicircle or "moon" on a black matte table surface. The experimental set-up is presented in Figure 1. A video camera above the surface provides a video image that is processed by a color-based recognition and tracking system (Smart – Panlab, Barcelona Spain) that generates a time ordered sequence of the contacts that occur between objects that is subsequently processed for event analysis (below). The simultaneous narration of the ongoing events is processed by a commercial speech-to-text system (IBM ViaVoice™). Speech and vision data were acquired and then processed off-line yielding a data set of matched sentence – scene pairs that were provided as input to the structure mapping model. A total of ~300 <sentence, scene> pairs were tested in the following experiments.

Figure 1. Experimental set-up for generating visual scene and speech input for the system. A video camera views events that are generated by the human operator, who at the same time narrates the ongoing events.

2 Visual Scenes and analysis

For a given video sequence the visual scene analysis generates the corresponding event description in the format *event(agent, object, recipient).*

2.1 Single Event Labeling

Events are defined in terms of contacts between elements. A contact is defined in terms of the time at which it occurred, the agent, object, and duration of the contact. The agent is determined as the element that had a larger relative velocity towards the other element involved in the contact. Based on these parameters of contact, scene events are recognized as follows:

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

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

Take(agent, object, source): Multiple contacts, as the agent takes the object from the source. 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.

Give(agent, object, recipient): In this multiple contact event, the agent first takes the object, and then 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.

These event labeling templates form the basis for a template matching algorithm that labels events based on the contact list, similar to the spanning interval and event logic of Siskind [24].

2.2 Complex "Hierarchical" Events

The events described above are simple in the sense that there have no hierarchical structure. This imposes serious limitations on the syntactic complexity of the corresponding sentences ([11, 18]). The sentence "The block that pushed the moon was touched by the triangle" illustrates a complex event that exemplifies this issue. The corresponding compound event will be recognized and represented as a pair of temporally successive simple event descriptions, in this case: *push(block, moon)*, and *touch(triangle, block)*. The "block" serves as the link that connects these two simple events in order to form a complex hierarchical event.

3. Structure mapping for language learning

The mapping of sentence form onto meaning [12] takes place at two distinct levels: Words are associated with individual components of event descriptions, and grammatical structure is associated with functional roles within scene events. The first level has been addressed by Siskind [23], Roy & Pentland [21] and Steels [25] and we treat it here in a relatively simple but effective manner. Our principle interest lies more in the second level of mapping between scene and sentence structure.

3.1 Word Meaning

In the initial learning phases there is no influence of syntactic knowledge and the word-referent associations are stored in the WordToReferent matrix (Eqn 1) by associating every word with every referent in the current scene ($\alpha = 1$), exploiting the cross-situational regularity [23] that a given word will have a higher coincidence with referent to which it refers than with other referents. This initial word learning contributes to learning the mapping between sentence and scene structure (Eqn. 4, 5 & 6 below). Then, knowledge of the syntactic structure, encoded in FormToMeaning can be used to identify the appropriate referent (in the SEA) for a given word (in the OCA), corresponding to a zero value of α in Eqn. 1. In this "syntactic bootstrapping" for the new word "gugle," for example, syntactic knowledge of Agent-Event-Object structure of the sentence "John pushed the gugle" can be used to assign "gugle" to the object of push.

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

3.2 Open vs Closed Class Word Categories

Our approach is based on the cross-linguistic observation that open class words (e.g. nouns, verbs, adjectives and adverbs) are assigned to their thematic roles based on word order and/or grammatical function words or morphemes [1]. Newborn infants are sensitive to the perceptual properties that distinguish these two categories [22], and in adults, these categories are processed by dissociable neurophysiological systems [2]. Similarly, artificial neural networks can also learn to make this function/content distinction [19]. Thus, for the speech input that is provided to the learning model open and closed class words are directed to separate processing streams that preserve their order and identity, as indicated in Figure 2.

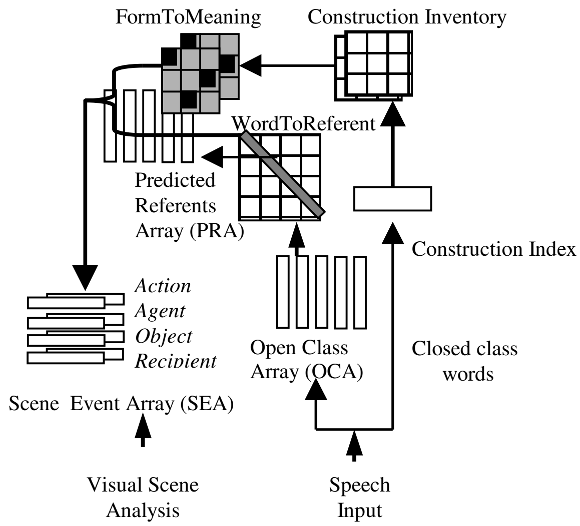

Figure 2. Structure-Mapping Architecture. Open class words in OCA are translated to Predicted Referents in the PRA via the WorldToReferent mapping. PRA elements are mapped onto their roles in the SEA by the FormToMeaning mapping, specific to each sentence type. This mapping is retrieved from Construction Inventory, via the ConstructionIndex that encodes the closed class words that characterize each sentence type.

3.3 Mapping Sentence to Meaning

In terms of the architecture in Figure 2, this mapping can be characterized in the following successive steps. First, words in the Open Class Array are decoded into their corresponding scene referents (via the WordToReferent mapping) to yield the Predicted Referents Array that contains the translated words while preserving their original order from the OCA (Eqn 2).

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

Next, each sentence type will correspond to a specific *form to meaning* mapping between the PRA and the SEA. encoded in the FormToMeaning array. The problem will be to retrieve for each sentence type, the appropriate corresponding FormToMeaning mapping. To solve this problem, we recall that each sentence type will have a unique constellation of closed class words and/or bound morphemes (Bates et al. 1982) that can be coded in a ConstructionIndex (Eqn.3) that forms a unique identifier for each sentence type. Thus, the appropriate FormToMeaning mapping for each sentence type can be indexed in ConstructionInventory by its corresponding ConstructionIndex.

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

The link between the ConstructionIndex and the corresponding FormToMeaning mapping is established as follows. As each new sentence is processed, we first reconstruct the specific FormToMeaning mapping for that sentence (Eqn 4), by mapping words to referents (in PRA) and referents to scene elements (in SEA). The resulting, FormToMeaningCurrent 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 FormToMeaningCurrent will depend on the quality of acquired word meanings in WordToReferent. Thus, syntactic learning requires a minimum baseline of semantic knowledge.

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

Given the FormToMeaningCurrent mapping for the current sentence, we can now associate it in the ConstructionInventory with the corresponding function word configuration or ConstructionIndex for that sentence, expressed in (Eqn 5).

$$\text{ConstructionInventory}(i,j) = \text{ConstructionInventory}(i,j) + \text{ConstructionIndex}(i) * \text{FormToMeaning-Current}(j) \qquad (5)$$

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

$$\text{FormToMeaning}(i) = \sum_{i=1}^{n} \text{ConstructionInventory}(i,j) * \text{ConstructinIndex}(j) \qquad (6)$$

To accommodate the dual scenes for complex events Eqns. 4-7 are instantiated twice each, to represent the two components of the dual scene. In the case of simple scenes, the second component of the dual scene representation is null.

We evaluate performance by using the WordToReferent and FormToMeaning 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 FormToMeaning transformation as described in Eqn. 7.

$$PSA(m,i) = PRA(k,i) * FormToMeaning(m,k) \qquad (7)$$

When learning has proceeded correctly, the predicted scene array (PSA) contents should match those of the scene event array (SEA) that is directly derived from input to the model. We then quantify performance error in terms of the number of mis-matches between PSA and SEA.

4. Experimental results

Hirsh-Pasek & Golinkof [14] indicate that children can use knowledge of word meaning to acquire a fixed SVO template around 18 months, and then expand this to non-canonical sentence forms around 24+ months. Tomasello [28] similarly indicates that fixed grammatical constructions will be used initially, and that these will then provide the basis for the development of more generalized constructions [12]. The following experiments attempt to follow this type of developmental progression.

4.1 Learning of Active Forms for Simple Events

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

For this experiment, 17 scene/sentence pairs were generated that employed the 5 different events, and narrations in the active voice, corresponding to the grammatical forms 1 and 2. The model was trained for 32 passes through the 17 scene/sentence pairs for a total of 544 scene/sentence pairs (Exp A in Figure 3). During the first 200 scene/sentence pair trials, α in Eqn. 1 was 1 (i.e. no syntactic bootstrapping before syntax is adquired), and thereafter it was 0. This was necessary in order to avoid the random effect of syntactic knowledge on semantic learning in the initial learning stages. The trained system displayed error free performance for all 17 sentences, and generalization to new sentences that had not previously been tested.

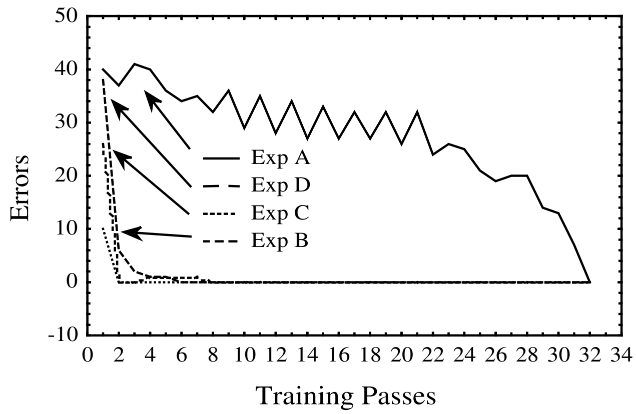

Figure 3. Evolution of interpretation errors during learning. The progressive reduction of interpretation errors in Exp A corresponds to the initial acquisition of the lexicon. In subsequent experiments exploiting this fixed lexicon, further interpretation errors are significantly reduced despite the increase of syntactic complexity in Exps B, C and D.

4.2 Passive forms

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

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

Seventeen new scene/sentence pairs were generated with active and passive grammatical forms for the narration. Within 3 training passes through the 17 sentences (51 scene/sentence pairs), error free performance was achieved, with confirmation of error free generalization to new untrained sentences of these types (Exp B in Figure 3). The rapid learning indicates the importance of lexicon in establishing the form to meaning mapping for the grammatical constructions.

4.3 Relative forms for Complex Events

Here we consider complex scenes narrated by sentences with relative clauses. Eleven complex scene/sentence pairs were generated with narration corresponding to the grammatical forms indicated in 5 – 10:

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

After presentation of 88 scene/sentence pairs, the model performed without error for these 6 grammatical forms, and displayed error-free generalization to new sentences that had not been used during the training for all six grammatical forms (Exp C in Figure 3).

4.4 Combined Test with and Without Lexicon

A total of 27 scene/sentence pairs, used in Experiments B and C, were employed that exercised the ensemble of grammatical forms 1 – 10 using the learned WordToReferent mappings. After exposure to 162 scene/sentence pairs the model performed and generalized without error (Exp D in Figure 3). When this combined test was performed without the pre-learned lexical mappings in WordToReferent, the system failed to converge, illustrating the advantage of following the developmental progression from lexicon to simple to complex grammatical structure.

4.5 Effects of "Syntactic" Knowledge on New Word Learning

One of the great strengths of a syntactic system is that the acquired rules are not only appropriate for processing previously learned sentences, but they can also be applied to new sentences in an open-ended generalization. In this case, the use of syntactic knowledge should aid in the acquisition of new word meanings, by allowing children to identify the thematic role of a new word, based on its configuration within a sentence [14]. Furthermore, as the syntactic knowledge is progressively acquired we should see that new word meanings can be acquired with increasing efficacy.

We thus exposed the model to new verbs at three distinct points (early, middle and late) in its progressive development of semantic and syntactic knowledge. The early period corresponds to the outset of language learning. The middle point corresponds to exposure to 400 sentences, and the late period corresponds to exposure to 4000 sentences. The training sentences were made up of active and passive two-argument verb sentences. At each of these three developmental periods we examined the strength of learning for the new verb after three different periods of exposure to this new verb (Period 1 = 33 exposures, Period 2 = 66 and Period 3 = 165 exposures) in a mixture of active and passive sentences.

The results of this experiment are presented in Figure 4 where we can observe that the learning strength increases with the duration of exposure for all three cases of syntactic acquisition. However, the main point is that for the condition in which the syntax has been most completely acquired (Late) there is a significant advantage in the learning strength for the new verb.

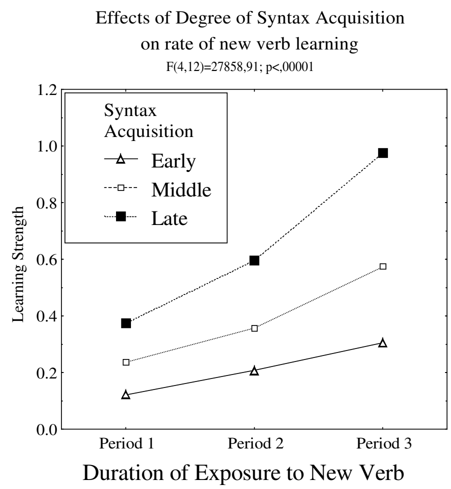

Figure 4. Effects of syntax on new verb learning. Learning strength for a new verb is evaluated at three different training exposure periods for models that have different levels of syntax acquisition. The most effective learning is observed for the models that have a greater degree of syntax acquisition (from [7]).

4.6 Some Scaling Issues

A small lexicon and construction inventory are used to illustrate the system behavior. Based on the independant representation formats, the architecture should scale well. The has now been tested with a larger lexicon, and has learned over 35 grammatical constructions, presented in the Appendix. The system should extend to all languages in which sentence to meaning mapping is encoded by word order and/or grammatical marking [1]. In the current study, deliberate human event production yielded essentially perfect recognition, though the learning model is relatively robust [7] to elevated scene error rates.

5. Discussion and Conclusion

The current results, along with related results in the developing domain of robot cognition (see [21, 24, 25]) can be considered in the context of the symbol grounding problem posed by Harnad [13]. The problem concerns how symbols can become functionally linked to their real world referents and meanings. Harnad proposed a solution that required symbols to be grounded by sensor-based "iconic" representations of distal objects, and by learning-based "categorical" representations that capture invariant properties. These representations could then be used in a combinatorial manner to construct composite representations. In this context, Harnad proposed a hybrid architecture in which invariant categorical relations would be learned by a connectionist system providing the grounding for the partner symbol manipulation system.

The current system can be considered in this context. Lexical elements corresponding to nouns in the Open Class Array (symbols) are "grounded" in their meaning/referent in the Scene Event Array based on object recognition performed by the visual scene analysis. These relations are encoded in learned connections in the WordToReferent matrix. Interestingly, for lexical elements corresponding to verbs in the OCA, the corresponding processing in the Visual Scene Analysis requires categorical recognition of different event/verb types based on the characteristics of the elementary contact event (Section 2). Thus, as illustrated by Siskind [24], the characterization of events in terms of force dynamic features such as contact yields a robust and highly invariant categorization capability.

Finally, the symbol grounding problem can be considered both at the lexical and the phrasal levels. At the phrasal level, grammatical constructions are grounded in their event structure representations via learned categorical relations in the FormToMeaning matrix encoded in the ConstructionInventory. This is consistent with the idea that both at the lexical and phrasal levels, iconic representations initiate the grounding process and are later generalized to categorical representations (see Tomasello [28]).

The current study demonstrates (1) that the perceptual primitive of contact (available to infants at 5 months), can be used to perform event description in a manner that is similar to but significantly simpler than Siskind [24], (2) that a novel implementation of principles from construction grammar can be used to map sentence form to these meanings together in an integrated system, (3) that relative clauses can be processed in a manner that is similar to, but requires less specific machinery (e.g. no stack) than that in Miikkulainen [18], and finally (4) that the resulting system displays robust acquisition behavior that reproduces certain observations from developmental studies with very modest “innate” language specificity.

It is important to note that some processing capabilities must be presumed innate. In the generativist framework [3] a full-fledged Universal Grammar is considered innate. The innate component is highly predetermined, largely in response to the presumed indeterminacy of the language acquisition problem. In the current study, the "innate" component is orders of magnitude simpler than UG, and corresponds to a generalized structure-to-structure mapping capability [6-8].

Note that one could have taken the same approach by integrating Siskind's [24] full event system, and Miikkulainen's [18] embedded case-role system. Each of these however required significant architectural complexity to accomplish the full job. Again, the current goal was to identify minimal event recognition and form-to-meaning mapping capabilities that could be integrated into a coherent system that performs at the level of a human infant in the first years of development when the construction inventory is being built up. This forms the basis for the infant's subsequent ability to de- and re-compose these constructions in a truly compositional manner, a topic of future research.

Acknowledgments

Supported by the EuroCores OMLL project, the French ACI Integrative and Computational Neuroscience Project, and the HFSP MCILA Project.

References

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

[2] 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

[3] Chomsky N. (1995) The Minimalist Program. MIT

[4] Chang NC, Maia TV (2001) Grounded learning of grammatical constructions, *AAAI Spring Symp. On Learning Grounded Representations*, Stanford CA.

[5] Crangle C. & Suppes P. (1994) Language and Learning for Robots, CSLI lecture notes: no. 41, Stanford.

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

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

[8] 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*.

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

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

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

[12] Goldberg A (1995) *Constructions*. U Chicago Press, Chicago and London.
[13] Harnad, S. (1990) The Symbol Grounding Problem. Physica D 42: 335-346.
[14] Hirsh-Pasek K, Golinkof RM (1996) *The origins of grammar: evidence from early language comprehension.* MIT Press, Boston.
[15] Kotovsky L, Baillargeon R, The development of calibration-based reasoning about collision events in young infants. 1998, *Cognition*, 67, 311-351
[16] Langacker, R. (1991). *Foundations of Cognitive Grammar. Practical Applications, Volume 2*. Stanford University Press, Stanford.
[17] 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
[18] Miikkulainen R (1996) Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science*, 20:47-73.
[19] 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 (Eds) *Signal to syntax*, Lawrence Erlbaum, Mahwah NJ, USA.
[20] Pollack JB (1990) Recursive distributed representations. Artificial Intelligence, 46:77-105.
[21] Roy D, Pentland A (2002). Learning Words from Sights and Sounds: A Computational Model. *Cognitive Science*, 26(1), 113-146.
[22] 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.
[23] Siskind JM (1996) A computational study of cross-situational techniques for learning word-to-meaning mappings, *Cognition* (61) 39-91.
[24] 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
[25] Steels, L. (2001) Language Games for Autonomous Robots. *IEEE Intelligent Systems*, vol. 16, nr. 5, pp. 16-22, New York: IEEE Press.
[26] Stolcke A, Omohundro SM (1994) Inducing probablistic grammars by Bayseian model merging/ In *Grammatical Inference and Applications: Proc. 2nd Intl. Colloq. On Grammatical Inference*, Springer Verlag.
[27] Talmy L (1988) Force dynamics in language and cognition. *Cognitive Science*, 10(2) 117-149.
[28] Tomasello M (1999) The item-based nature of children's early syntactic development, *Trends in Cognitive Science*, 4(4):156-163

Appendix: Sentence and scene descriptions

The <sentence, meaning> pairs for training and testing are constructed from the following templates. The lexicon consists of 5 nouns (cylinder, moon, block, cat, dog), 5 verbs (touch, push, take, give, say), and 8 function words (to, by, from, was, that, it, itself, and)[1].

[1] Possible scaling issues for WordToWorld mappings are not of concern here. If WordToWorld is well specified, then lexicon size has no influence on SentenceToWorld mapping.

A.1 Single event scenes

1. Agent verb object. (Active)
   Verb(agent, object)
2. Object was verbed by agent. (Passive)
   Verb(agent, object).
3. Agent verbed object to recipient. (Dative)
   Verb(agent, object, recipient)
4. Object was verbed to recipient by agent. (Dative passive)
   Action1(agent1, object2, recipient3).
5. Agent1 action1 recipient3 object2.
   Verb(agent, object, recipient).

A.2 Double event relatives

6. Agent1 that verb1ed object2 verb2ed object3. (Relative agent).
   Action1(agent1,object2), Action2(agent1,object3)
7. Object3 was action2ed by agent1 that action1ed object2. (Relative object).
   Action1(agent1,object2), Action2(agent1,object3)
8. Agent1 that action21ed object2 was action22ed by agent3
   Action1(agent1,object2), Action2(agent3,object1)
9. Agent3 action2ed object1 that action1ed object2
   Action1(agent1,object2), Action2(agent3,object1)
10. Obj2 that was action1ed by agent1 action2ed obj3
    Action1(agent1,object2), Action2(agent2,object3)
11. Obj3 was act2d by agent2 that was act1d by agent1
    Action1(agent1,object2), Action2(agent2,object3)
12. Obj2 that was action1ed by agent1 was action2ed by ag3
    Action1(agent1,object2), Action2(agent3,object2)
13. ag3 act22ed obj2 that was act21ed by ag1
    Action21(agent1,object2), Action22(agent3,object2)
14. Ag1 that act1ed obj2 act2ed obj3 to recip4
    Action1(agent1,object2), Action2(agent1,object3,recipient4)
15. Obj3 was act32ed to recip4 by ag1 that act21ed obj2
    Action1(agent1,object2), Action2(agent1,object3,recipient4)
16. Agent1 that action1ed object2 was action2ed to recip4 by ag3
    Action1(agent1,object2), Action2(agent3,object1,recipient4)
17. Ag3 act2ed obj4 to recip1 that act1ed obj2
    Action1(agent1,object2), Action2(agent3,object4,recipient1)
18. Obj4 was act2ed from ag3 to recip1 that act1ed obj2
    Action1(agent1,object2), Action2(agent3,object4,recipient1)
19. Obj2 that was act1ed by ag1 act2ed obj3 to recip4
    Action1(agent1,object2), Action2(agent2,object3,recipient4)
20. Ag3 act2ed ob4 to rec2 that was act1ed by ag1

Action1(agent1,object2), Action2(agent3,object4,recipient2)

21. Ag1 that act1ed obj2 to rec3 act2ed obj4
    Action1(agent1,object2,recipient3), Action2(agent1,object4)
22. Obj4 was act2ed by ag1 that act1ed ob2 to rec3
    Action1(agent1,object2,recipient3), Action2(agent1,object4)

23. Ag4 act2ed ob1 that act1ed ob2 to rec3
    Action1(agent1,object2,recipient3), Action2(agent4,object1)

24. Ob1 that act1ed ob2 to rec3 was act2ed by ag4
    Action1(agent1,object2,recipient3), Action2(agent4,object1)
25. Ag2 that was act1ed by ag1 to rec3 act2ed ob4
    Action1(agent1,object2,recipient3), Action2(agent2,object4)
26. Ag4 act2ed obj2 that was act1ed by ag1 to rec3
    Action1(agent1,object2,recipient3), Action2(agent4,object2)

A.3 Dual event Conjoined

27. Agent1 action1 object1 and object2. (Active conjoined object)
    Action1(agent1, object1),
    Action1(agent1, object2)
28. Agent1 and agent3 action1ed object2. (Active conjoined agent)
    Action1(agent1, object2),
    Action1(agent3, object2)
29. Agent1 action1ed object2 and action2 object3. (Conjoined)
    Action1(agent1, object2),
    Action2(agent1, object3)

A.4 Dual Event Reflexive

30. Agent1 action1r that agent2 action2ed object3. (Simple reflexive)
    Action1r[2](agent1),
    Action2(agent2, object3).
31. Agent1 action1ed itself. (Simple active reflexive)
    Action1(agent1, agent1).
32. Agent1 action1r that agent2 action2ed itself. (Reflexive simple noun phrase).
    Action1r(agent1),
    Action2(agent2, agent2).
33. Agent1 action1r that agent2 action2ed it. (Pronoun simple noun phrase).
    Action1r(agent1),
    Action2(agent2, agent1).
34. Agent1 action1r that it action1ed object2.
    Action1r(agent1),

[2] Corresponds to reflexive verbs such as "said," or "believed."

Action2(agent1, object2).

35. Agent1 action1r that object3 was action2ed by agent2.
    Action1r(agent1),
    Action2(agent2, object3).
36. Agent1 action1r that agent2 action2ed object3 to recipient4.
    Action1r(agent1),
    Action2(agent2, object3, recipient4).
37. Agent1 action1r agent2 action2ed object3 to recipient4.
    Action1r(agent1),
    Action2(agent2, object3, recipient4).
38. Object2 object3 were action1ed to recipient4 by agent1.
    Action1(agent1, object2, recipient4), Action1(agent1, object3, recipient4)