Vision and Action in the Language-Ready Brain: From Mirror Neurons to SemRep

Michael A. Arbib[1,2,3] and JinYong Lee[1]
[1] Computer Science, [2] Neuroscience, and [3] USC Brain Project
University of Southern California, Los Angeles, CA 90089-2520, USA
arbib@usc.edu, jinyongl@usc.edu

Abstract. The general setting for our work is to locate language perception and production within the broader context of brain mechanisms for action and perception in general, modeling brain function in terms of the competition and cooperation of schemas. Particular emphasis is placed on mirror neurons – neurons active both for execution of a certain class of actions and for recognition of a (possibly broader) class of similar actions. We build on the early VISIONS model of schema-based computer analysis of static scenes to present SemRep, a graphical representation of dynamic visual scenes designed to support the generation of varied descriptions of episodes. Mechanisms for parsing and production of sentences are currently being implemented within Template Construction Grammar (TCG), a new form of construction grammar distinguished by its use of SemRep to express semantics.

Keywords: action, action recognition, brain mechanisms, competition and cooperation, construction grammar, dynamic visual scenes, language perception, language production, mirror neurons, visual perception, scene descriptions, schema theory, SemRep, vision

1. Introduction

The present section provides the background for the novel material of this paper: Section 2, which presents SemRep, a graphical representation of dynamic visual scenes designed to support the generation of varied descriptions of episodes; and Section 3, which presents Template Construction Grammar (TCG), the version of construction grammar in which we locate our current efforts to implement mechanisms for the parsing and production of sentences. We also summarize the Mirror System Hypothesis, an evolutionary framework for analyzing brain mechanisms of language perception and production which places particular emphasis on the role of mirror neurons. We briefly note that the brain may be modeled in terms of the competition and cooperation of schemas. Finally, we recall key features of the early VISIONS model of schema-based computer analysis of static scenes to provide background for the design of SemRep.

1.1 Schemas Which Compete and Cooperate

Vision is often seen as a process that classifies visual input, e.g., recognizing faces from photographs, or segmenting a scene and labeling the regions, or detecting characteristic patterns of motion in a videoclip. However, our approach to vision is concerned with its relevance to the ongoing behavior of an embodied agent be it frog, rat, monkey, human or robot [1, 2] – we view vision under the general rubric of *action-oriented perception*, as the “active organism” seeks from the world the information it needs to pursue its chosen course of action. A *perceptual schema* not only determines whether a given "domain of interaction" (an action-oriented generalization of the notion of object) is present in the environment but can also provide parameters concerning the current relationship of the organism with that domain. *Motor schemas* provide the control systems which can be coordinated to effect the wide variety of movement.

A *coordinated control program* is a schema assemblage which processes input via perceptual schemas and delivers its output via motor schemas, interweaving the activations of these schemas in accordance with the current task and sensory environment to mediate more complex behaviors [3]. A given action may be invoked in a wide variety of circumstances; a given perception may precede many courses of action. There is no one grand "apple schema" which links all "apple perception strategies" to "every action that involves an apple". Moreover, in the schema-theoretic approach, "apple perception" is not mere categorization – "this is an apple" – but may provide access to a range of parameters relevant to interaction with the apple at hand.

1.2 The VISIONS System

An early example of schema-based interpretation for visual scene analysis in the VISIONS system [4]. However, it is *not* an action-oriented system, but rather deploys a set of perceptual schemas to label objects in a static visual scene. In VISIONS, there is no extraction of gist – rather, the gist is prespecified so that only those schemas are deployed relevant to recognizing a certain kind of scene (e.g., an outdoor scene with houses, trees, lawn, etc.). Low-level processes take an image of such an outdoor visual scene and extract and builds a representation in the *intermediate database* – including contours and surfaces tagged with features such as color, texture, shape, size and location. An important point is that the segmentation of the scene in the intermediate database is based not only on bottom-up input (data-driven) but also on top-down hypotheses (e.g., that a large region may correspond to two objects, and thus should be resegmented).

VISIONS applies perceptual schemas across the whole intermediate representation to form confidence values for the presence of objects like houses, walls and trees. The schemas are stored in LTM (long-term memory), while the state of interpretation of the particular scene unfolds in STM (short-term or working memory) as a network of schema instances which link parameterized copies of schemas to specific portions of the image to represent aspects of the scene of continuing relevance.

Interpretation of a novel scene starts with the data-driven instantiation of several schemas (e.g., a certain range of color and texture might cue an instance of the foliage schema for a certain region of the image). When a schema instance is activated, it is linked with an associated area of the image and an associated set of local variables. Each schema instance in STM has an associated confidence level which changes on the basis of interactions with other units in STM. The STM network makes context explicit: each object represents a context for further processing. Thus, once several schema instances are active, they may instantiate others in a “hypothesis-driven” way (e.g., recognizing what appears to be a roof will activate an instance of the house schema to seek confirming evidence in the region below that of the putative roof). Ensuing computation is based on the competition and cooperation of concurrently active schema instances. Once a number of schema instances have been activated, the schema network is invoked to formulate hypotheses, set goals, and then iterate the process of adjusting the activity level of schemas linked to the image until a coherent interpretation of (part of) the scene is obtained. VISIONS uses *activation values* so that schema instances may compete and cooperate to determine which ones enter into the equilibrium schema analysis of a visual scene. (The HEARSAY speech understanding system [5] extends this into the time domain. In HEARSAY, entities at different levels – phonemes, words, phrases and sentences compete and cooperate to cover certain time periods of the auditory input in a consistent fashion. But in the end, what emerges is that single coherent symbolic representation.) Cooperation yields a pattern of "strengthened alliances" between mutually consistent schema instances that allows them to achieve high activity levels to constitute the overall solution of a problem. As a result of competition, instances which do not meet the evolving consensus lose activity, and thus are not part of this solution (though their continuing subthreshold activity may well affect later behavior). Successful instances of perceptual schemas become part of the current short-term model of the environment.

The classic VISIONS system had only a small number of schemas at its disposal, and so could afford to be lax about scheduling their application. However, for visual systems operating in a complex world, many schemas are potentially applicable, and many features of the environment are interpretable. In this case, “attention” – the scheduling of resources to process specific parts of the image in particular ways – becomes crucial. How this may be accomplished is described elsewhere [6], as is the way in which VISIONS may be extended to mediate action-oriented perception by an agent in continuous interaction with its environment [2].

1.3 From Visual Control of Grasping to Mirror Neurons

The minimal neuroanatomy of the brain of the macaque monkey and the human (or of mammals generally) that we need here is that the cerebral cortex can be divided into four lobes: the *occipital lobe* at the back (which includes primary visual cortex); the *parietal* lobe (moving up and forward from the occipital lobe); the *frontal lobe* and then moving back beneath frontal and parietal cortex, the *temporal lobe*. *Prefrontal cortex* is *at* the front of the frontal lobe, not *in* front of the frontal lobe. For the moment, we are particularly interested in three areas:

- Parietal area AIP, which is the anterior region within a fold of parietal cortex called the intra-parietal sulcus,
- A ventral region of premotor area called F5, and

- Inferotemporal cortex (IT), a region of the temporal lobe particularly associated with object recognition.

AIP and F5 anchor the cortical circuit in macaque which transforms visual information on intrinsic properties of an object into hand movements for grasping it. Discharge in most grasp-related F5 neurons correlates with an action rather than with the individual movements that form it so that one may relate F5 neurons to various *motor schemas* corresponding to the action associated with their discharge:

The FARS (Fagg-Arbib-Rizzolatti-Sakata) model [7] addresses key data on F5 and AIP from the labs of Giacomo Rizzolatti in Parma and Hideo Sakata in Tokyo, respectively. In the FARS model, area cIPS (another parietal area – the details do not matter for this exposition) provides visual input to parietal area AIP concerning the position and orientation of the object's surfaces. AIP then extracts the *affordances* the object offers for grasping (i.e., the visually grounded encoding of "motor opportunities" for grasping the object, rather than its classification [8]). The basic pathway AIP → F5 → F1 (primary motor cortex) of the FARS model then transforms the (neural code for) the affordance into the coding for the appropriate motor schema in F5 and thence to the appropriate detailed descending motor control signals (F1).

Going beyond the empirical data then available, FARS [7] stressed that there may be several ways to grasp an object and thus hypothesized (a) that object recognition (mediated by IT) can affect the computation of working memory, task constraints and instruction stimuli in various parts of prefrontal cortex (PFC), and (b) that strong connections from PFC can bias the selection in the AIP → F5 pathway of which grasp to execute. The two major paths from visual cortex via parietal cortex (e.g., AIP) and inferotemporal cortex (e.g., IT) are labeled as the *dorsal* and *ventral* paths, respectively. The dorsal path is concerned with the "how" or parameterization of action, while the ventral path encodes the "what" or knowledge of action, appropriate to planning a course of action rather than the fine details of its execution.

To proceed, we must note the discovery of a very significant subset of the F5 neurons related to grasping – the *mirror neurons*. These neurons are active not only when the monkey executes a specific hand action but also when it observes a human or other monkey carrying out a similar action. These neurons constitute the "mirror system for grasping" in the monkey and we say that these neurons provide the neural code for matching execution and observation of hand movements. (By contrast, the *canonical* neurons – which are the F5 neurons that actually get modeled in FARS – are active for execution but not for observation.) A mirror *system* for a class X of actions is a region of the brain that, compared with other situations, becomes more active both when actions from class X are observed and when actions from class X are executed. Mirror neurons exist for a range of actions in the macaque monkey, and brain imaging experiments have demonstrated a mirror system for grasping in the human, but we have no single neuron studies proving the reasonable hypothesis that the human mirror *system* for grasping contains mirror neurons for specific grasps. In work not reported here, we are extending our models of the mirror system [9, 10] from hand movements to action recognition more generally. Our prior models are based on neural networks for recognition of trajectory of the hand relative to an object. They use an object-centered coordinate system to recognize whether the hand is on track to perform a particular action upon the object, which may explain data in [11].

1.4 From Mirror Neurons to the Mirror System Hypothesis

Area F5 in the macaque is homologous to area 44 in the human, part of Broca's area, an area normally associated with speech production. Yet this area in humans contains a mirror system to grasping. These data led Arbib & Rizzolatti [12] to develop the *Mirror-System Hypothesis* – Language evolved from a basic mechanism *not* originally related to communication: the *mirror system for grasping* with its capacity to generate *and* recognize a set of actions. More specifically, human Broca's area contains a mirror system for grasping which is homologous to the F5 mirror system of macaque, and this provides the evolutionary basis for *language parity* – namely that an utterance means roughly the same for both speaker and hearer.

This provides a neurobiological "missing link" for the hypothesis that communication based on manual gesture preceded speech in language evolution.

Arbib [13] has amplified the original account of Rizzolatti and Arbib to hypothesize seven stages in the evolution of language. Rather than offer details here, we simply note the synthesis of ideas on the dorsal and ventral pathways with the concept of mirror neurons and schema assemblages provided by [14].

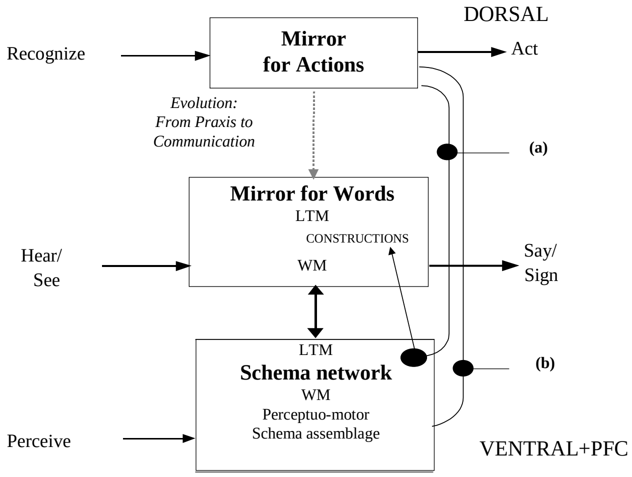

Fig. 1. Words link to schemas, not directly to the dorsal path for actions (from [14]).

Saussure [15] distinguishes the *Signifier* from the *Signified* (or words from concepts), but then highlights the "Sign" as combining these with the linkage between them. Our action-oriented view is that the basic concepts are realized as the perceptual and motor schemas of an organism acting in its world, and that that there is no direct labeling of one word for one concept. Rather, the linkage is many-to-one, competitive and contextual, so that appropriate words to express a schema may vary from occasion to occasion, both because of the assemblage in which the schema instance is currently embedded, and because of the state of the current discourse. Let us diagram this in a way which makes contact with all that has gone before. The lower 2 boxes of Figure 1 correspond to words and concepts, but we now make explicit, following the Mirror System Hypothesis, that we postulate that a mirror system for phonological expression ("words") evolved atop the mirror system for grasping to serve communication integrating hand, face and voice. We also postulate that the concepts – for diverse actions, objects, attributes and abstractions – are represented by a network of concepts stored in LTM, with our current "conceptual content" formed as an assemblage of schema instances in Working Memory (WM – compare the STM of VISIONS). Analogously, the Mirror for Words contains a network of word forms in LTM and keeps track of the current utterance in its own working memory.

The perhaps surprising aspect of the conceptual model shown here is that the arrow linking the "Mirror for Actions" to the "Mirror for Words" expresses an evolutionary relationship, not a flow of data. Rather than directly linking the dorsal action representation to the dorsal representation of phonological form, we have two relationships between the dorsal pathway for the Mirror for Actions and the schema networks and assemblages of the ventral pathway and prefrontal cortex (PFC). The rightmost path in Figure 1 corresponds to the paths in FARS whereby IT and PFC can affect the pattern of dorsal control of action. The path just to the left of this shows that the dorsal representation of actions can only be linked to verbs via schemas.

Rather than pursuing the study of brain mechanisms further, we work within the framework provided by [6] to ask the following: "If we extend our interest in vision from the recognition of the disposition of objects in static scenes to the relations between agents, objects and actions dynamic visual scenes, what sort of representations are appropriate to interface the visual and language systems?"

2 SemRep: A Semantic Representation for Dynamic Visual Scenes

SemRep is a hierarchical graph-like representation of a visual scene, whether static or dynamically extended over time (an episode). A SemRep graph structure represents the semantics of *some* of the cognitively salient elements of the scene. We see SemRep as an extension of the schema assemblages generated by the VISIONS system, but with the crucial addition of actions and of extension in time. Only cognitively important events are encoded into SemRep while others are simply discarded or absorbed into other entities. The same scene can have many different SemReps, depending on the current task and on the history of attention. A prime motivation is to ensure that this representation be usable to produce sentences that describe the scene, allowing SemRep to bridge between vision and language.

The structure of SemRep does not have to follow the actual changes of an event of interest, but may focus on "conceptually significant changes" – a crucial difference from a sensorimotor representation, where motor control requires continual tracking of task-related parameters. For example, an event describable by the sentence "Jack kicks a ball into the net" actually covers several time periods: [Jack's foot swings] → [Jack's foot hits a ball] → [the ball flies] → [the ball gets into the net]. Note that [Jack's foot swings] and [Jack's foot hits a ball] are combined into [Jack kicks a ball], and [the ball flies] is omitted. This taps into a schema network, which can use stored knowledge to "unpack" items of SemRep when necessary. On the other hand, a Gricean convention makes it unlikely that SemRep will include details that can be retrieved in this way, or details that are already known to speaker and hearer.

The same principle is applied to the topology of SemRep entities. The arrangement of conceptual entities and their connections might or might not follow that of the actual images and objects. A description “a man without an arm”, for example, does not exactly match an actual object setting since it encodes the conceptual entity of an arm which is missing in the actual image. This relates to the previous point: one may need to include what is not in the image to block standard inferences in cases where they are inappropriate. This is akin the notion of inheritance in semantic networks.

Similarly, an event or entity with higher cognitive importance – or “discourse importance”, what the speaker wishes to emphasize for the hearer – will be assigned to a higher level in the hierarchy independently of the methodology by which the entity is specified. For instance, even if the vision system had specified MAN first and this led to the zooming in on the face, FACE might be ranked higher in SemRep than MAN if the former is currently of greater interest.

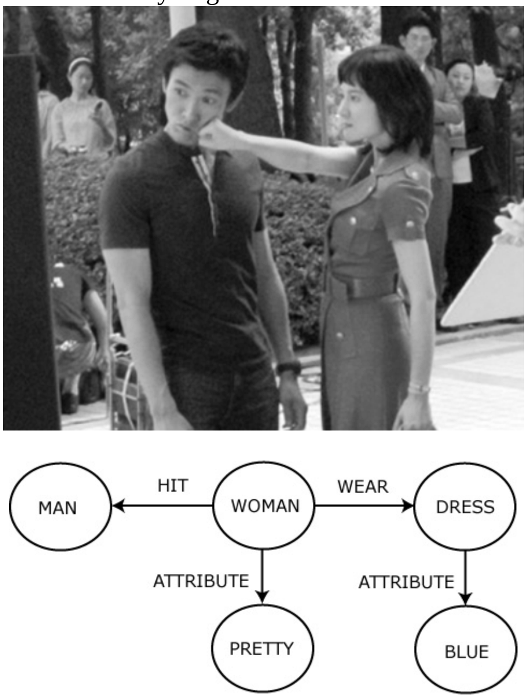

Fig. 2: Top: A picture of a woman hitting a man (original image from “Invisible Man Jangsu Choi”, Korean Broadcasting System). Bottom: A SemRep graph that could be generated for the picture. This might yield the sentence “A pretty woman in blue hits a man.”

In order to encode the various conceptual entities and their relationships, SemRep structure takes the form of a graph structure. The two major elements of a SemRep graph are ‘node’ and ‘relation (directed edge)’. Agents and various types of objects are usually represented as nodes, but we also use nodes to represent attributes. Relationships between nodes include actions linking agent and patient, spatial configuration, possessive relationship, movement direction or pointer which indicates the semantically identical node are represented as relations, as well as the relation between a node and its attributes. As mentioned above, a vision system can be one of the systems that create SemRep structure by imposing nodes and relations upon a visual image (or "videoclip"). An area interesting enough to capture attention is linked to a node (or a relation if an action is happening in that area) and then relations are specified among the found nodes, presumably by shifting attention. While most types of node and some types of relation – such as spatial, possessive, attributive relations – are established by static (spatial) analysis, action relations require dynamic (spatio-temporal) analysis.

Both nodes and relations may be attached to more detailed semantic descriptions defined as "conceptual structures". The properties of a recognized object are attached to a node for the object, and the semantics of an action are attached to an action relation. The attached concepts will later be translated into words by the language system. A relation includes the sets it relates and so a verb is not just a label for an action but incorporates restrictions on its slot fillers. However, the SemRep graph is not labeled with words but with more abstract descriptors, allowing the same graph to be expressed in multiple ways within a given language. Thus the concept YOUNG FEMALE could be translated into 'girl', 'woman' or even 'kid' and the action concept HITTING WITH HAND could be translated into 'hit', 'punch' or 'slap'. Again, the configuration where object A is placed vertically higher than B can be expressed as "A is above B", "B is below A", "A is on B", etc.

The action concept HIT may involve properties such as VIOLENT MOTION, BODY CONTACT, and CAUSING PAIN, and these properties implicitly show that the encoded concept describes an action. However, some of these processes may be directly perceptual (i.e., generated directly by the visual system) while others may be more inferential. It might be claimed [16] that mirror neurons will link action recognition to our own experience, so CAUSING PAIN might be perceived "directly", while the woman's ANGER might either be perceived directly or be more inferential.

Thus we view SemRep as providing a graphical structure which encompasses one analysis which captures a subset of the agents, objects, actions and relationships that may be present in a given (temporally extended) visual scene. Nodes in SemRep may also be given a *significance* value which expresses the importance of a particular aspect of the scene. Thus the same scene may be described by "John loves Mary" or "Mary is loved by John" depending on whether the focus (higher significance value) is given to John or Mary, respectively.

3. Template Construction Grammar (TCG)

Where many linguists operate within the framework of generative grammar (e.g., [17]), we work within the framework of *construction grammar* (e.g., [18, 19]). *Constructions* are form-meaning pairings which serve as basic building blocks for grammatical structure – each provides a detailed account of the pairing of a particular syntactic pattern with a particular semantic pattern, including phrase structures, idioms, words and even morphemes. By contrast, in generative grammar, meaning is claimed to be derived from the systematic combination of lexical items and the functional differences between the patterns that constructions capture are largely ignored.

Generative grammar distinguishes the lexicon from the grammar, which is seen as having three separate components – phonological, syntactic and semantic – with linking rules to map information from one component onto another. The rules of grammar are said to operate autonomously within each component, and any “rule breaking” within a particular language is restricted to idiosyncrasies captured within the lexicon. But what of idiomatic expressions like *kick the bucket*, *shoot the breeze, take the bull by the horns* or *climb the wall*? Should we consider their meanings as a supplement to the general rules of the syntactic and semantic components and their linking rules? Instead of this, Fillmore, Kay & O’Connor [20] suggested that the tools they used in analyzing idioms could form the basis for *construction grammar* as a new model of grammatical organization, with constructions ranging from lexical items to idioms to rules of quite general applicability [18]. Many linguists have teased out the rule-governed and productive linguistic behaviors specific to each family of constructions. Constructions, like items in the lexicon, cut across the separate components of generative grammar to combine syntactic, semantic and even in some cases phonological information. The idea of construction grammar is thus to abandon the search for separate rule systems within syntactic, semantic and phonological components and instead base the whole of grammar on the “cross-cutting” properties of constructions.

Going beyond this “intra-linguistic” analysis, we suggest that “vision constructions” may synergize with “grammar constructions” in structuring the analysis of a scene in relation to the demands of scene description [6] in a way which ties naturally to our discussion of VISIONS. We argue that the approach to language via a large but finite inventory of constructions coheres well with the notion of a large but finite inventory of “scene schemas” for visual analysis. Each constituent which expands a “slot” within a scene schema or verbal construction may be seen as a hierarchical structure in which extended attention to a given component of the scene extends the complexity of the constituents in the corresponding part of parse tree of a sentence. This enforces the view that visual scene analysis must encompass a wide variety of basic “schema networks” – more or less abstract SemReps in the conceptualization of the previous sentence – in the system of high-level vision, akin to those relating *sky* and *roof*, or *roof*, *house* and *wall* in the VISIONS system. Of course, we do not claim that all sentences are limited to descriptions of, or questions about, visual scenes, but we do suggest that understanding such descriptions and questions can ground an understanding of a wide range of language phenomena.

We are currently implementing parsing and production systems for our own version of construction grammar, Template Construction Grammar (TCG). TCG adopts two major policies of conventional construction grammar (CG): each construction specifies the mapping between form and meaning, and the systematic combination of constructions yields the whole grammatical structure. However, in TCG, the meaning of an utterance is given as a SemRep graph (with suitable extensions to be provided in further work). A SemRep may correspond to one or more sentences, basically by covering the relevant portion of the given SemRep with a set of “small” subgraphs, where each is chosen such that a construction is available which expresses that subgraph in the given language. Figure 3 shows a construction defined in TCG, exemplifying the links that indicate which part of a "SemRep template" connect to which slot in a text form. Each construction encodes the specification of what can be mapped to which text/slot, and the mapping is assumed to be bidirectional – it can be used in production of a sentence as well as for comprehension. Most other computational approaches to CG, such as Fluid Construction Grammar (FCG) [21], are based on the use of predicate logic rather than graphs as the basis for constructions.

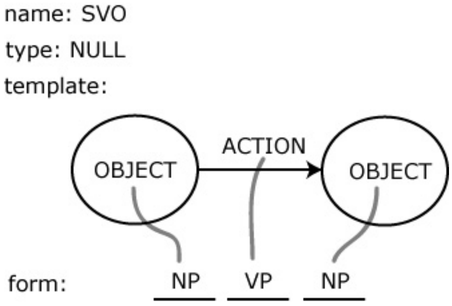

Fig. 3. An example '[subject] [verb] [object]' construction (a very general construction) in TCG. The template is an "abstract" SemRep, i.e., a graph like a SemRep but with either generic or (not in this case) specific labels on the edges and nodes, with each linked to a text or an empty slot for which there may be restrictions as to what can serve as slot fillers.

In production mode, the template acts to match constraints for selecting proper constructions by being superimposed on the SemRep graph that is going to be expressed in words. The semantic constraint of each construction is considered to be encoded in the template since the template also contains concepts as well as the topology of a SemRep graph. In comprehension mode, the template provides a frame where the interpreted meaning builds up as parsing progresses. The details of the interpreted SemRep graph are filled with the meaning of the constructions found by matching with the currently processed text (or word) one by one. Originally, form of each construction has to be a series of phonemes that would be combined into words, but it is assumed that these phonemes are already properly perceived and processed, and the correct words are given in a text form.

As mentioned above, the template is an (abstract) SemRep graph. The matching process in production mode is done by comparing the template of a construction to the given SemRep graph. The contents (given as the attached concepts) of nodes and relations and the connective structure of nodes and relations is considered in the process. The construction with the most 'similar' template will be chosen over other constructions, though provision must be made for backtracking. Note, too, that the similarity might be to a subgraph bottom up or a set of high-level nodes top-down – choices compete and cooperate till a spanning structure is formed. "Similarity" for the attached concepts is decided, for example, by how many common attributes they share – SemRep includes (though we have omitted the details from this paper) the ability to capture concepts by the superimposed distributed representation of attributes. Again, similarity for the structure of the template is decided by how close the topology of the template is to the given SemRep graph – the number of nodes and relations has to be matched as well as the connections between them.

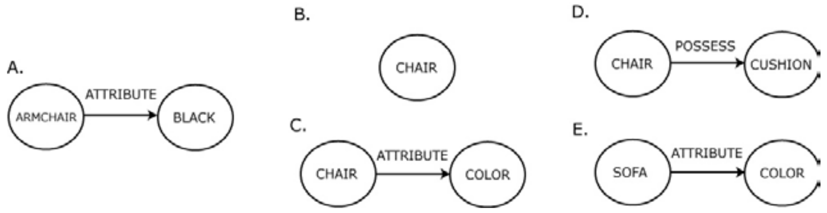

Fig. 4: SemRep graph A represents a 'black armchair'. Graphs B and C are "similar" to graph A but D and E are not.

Embedded structure is another topological feature to be considered. Matching requires that the template of a construction is a "subset" of the given SemRep graph. In other words, the template should not be more specific than the graph being compared. This rule applies to both concepts and topology. For example, in Figure 4 graph C is an appropriate match to graph A since ARMCHAIR is a kind of CHAIR and BLACK is a kind of COLOR and the topology is the same as that of graph A. Graph B is also appropriate because the topology is less specific. Graph D is inappropriate since the relations (ATTRIBUTE and POSSESSION) do not match each other; and Graph E is inappropriate since SOFA is a more detailed concept than ARMCHAIR. Among the appropriate graphs B and C, graph C will win over the competition because it is more similar to graph A than is B. If there were a graph identical to graph C except that it had BLACK node instead of COLOR, then this graph would have been the winner.

In the current version of TCG, the input text is assumed to be preprocessed and segmented into morphemes at a level that corresponds to the construction repertoire. Matching text can be somewhat simpler than matching templates since in matching text there is no need to perform complex comparison of graph structures. This is not to minimize the various obstacles to comprehending a sentence offered by anaphor, ellipsis, and ambiguity in interpretation, etc. And consider idiomatic expressions. For example, the idiom "a piece of cake" might be processed with a single construction which has the whole text in its form and the semantic meaning of "being easy". But it also can be processed with one or more general constructions. Allowing constructions with more specific information to be selected provides one possible default (in this case, idiomatic constructions would win over general constructions) but the eventual system must provide mechanisms for broader context to settle the issue: in parsing "Would you like a piece of cake?", the idiomatic construction is inappropriate.

In order to apply constructions hierarchically, each construction is assigned a *type* which specifies a sort of grammatical category for the construction, but such a type is not the highly abstract syntactic category of generative grammar, but is more like an emergent categorization rule generated by the relationship between constructions in the repertoire. Each empty slot in the form of a construction indicates the type of the construction that should fill the slot.

When translating the given SemRep graph into a sentence, the graph would activate a number of constructions with matching templates. In TCG, the construction with the best-matching template will be selected and its form will be output as the translated text, but if the form has any empty slot, it should be filled first. An empty slot specifies not only the type of construction that is expected, but also indicates the area of SemRep that is going to be considered for comparison; each slot is linked to a pre-specified area in the template, and only an area of SemRep corresponding to that area is considered for finding matching constructions for the slot. The link between the template and form provides the form-meaning pairing of a construction in TCG.

Since constructions are bidirectional, the same set of constructions used in production of sentence are also used in comprehension. All of the matching (or activated) constructions are eligible for translation until further processing reveals ineligibility. As input text is read, it is compared to the forms of activated constructions and the constructions with unmatched forms are ruled out. Ambiguity may also be resolved based on contextual information, which is in this case is the translated SemRep graph. However, top-down influences in sentence comprehension are beyond the scope of the current version of TCG.

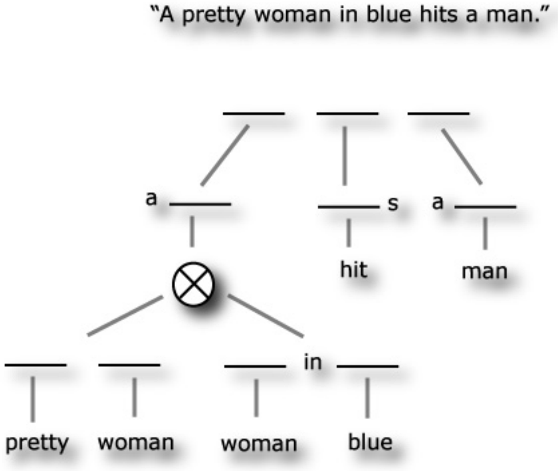

Fig. 5. The sentence "A pretty woman in blue hits a man" and the corresponding construction architecture. The language system would translate the SemRep graph in Figure 2 into the above sentence. During the process, constructions will be built into the hierarchical structure shown in the figure.

The type of the activated construction is also treated as input to the system and the matching mechanism is very similar to that for the text case, except that it is matched with the slot in the form rather than the text. For example, if an input sentence is given as "A big dog barks" then the first word "a" would activate at least two constructions, "a [adjective] [noun]" and "a [noun]" (or "[determinant] [adjective] [noun]" and "[determinant] [noun]" with "a" activating a construction of type [determinant]). Other configurations are possible, depending on the construction repertoire.

Given the activated constructions "a [adjective] [noun]" and "a [noun]", the next word "big" would activate a construction whose type is [adjective], ruling out the second construction due to mismatch of the construction type required in the slot.

Figure 5 shows one of the sentences that can be generated from the SemRep graph shown in Figure 2 and the resulted hierarchical build-up of constructions. Note that because of the multiple embedded structures in the WOMAN node, constructions for both "pretty woman" and "woman in blue" are present at the lowest level. These constructions are then combined into one expression "pretty woman in blue". The set of constructions might differ from those of other speakers to some extent. In that case, the constructions could be organized in a different structure and the hierarchy among constructions might change.

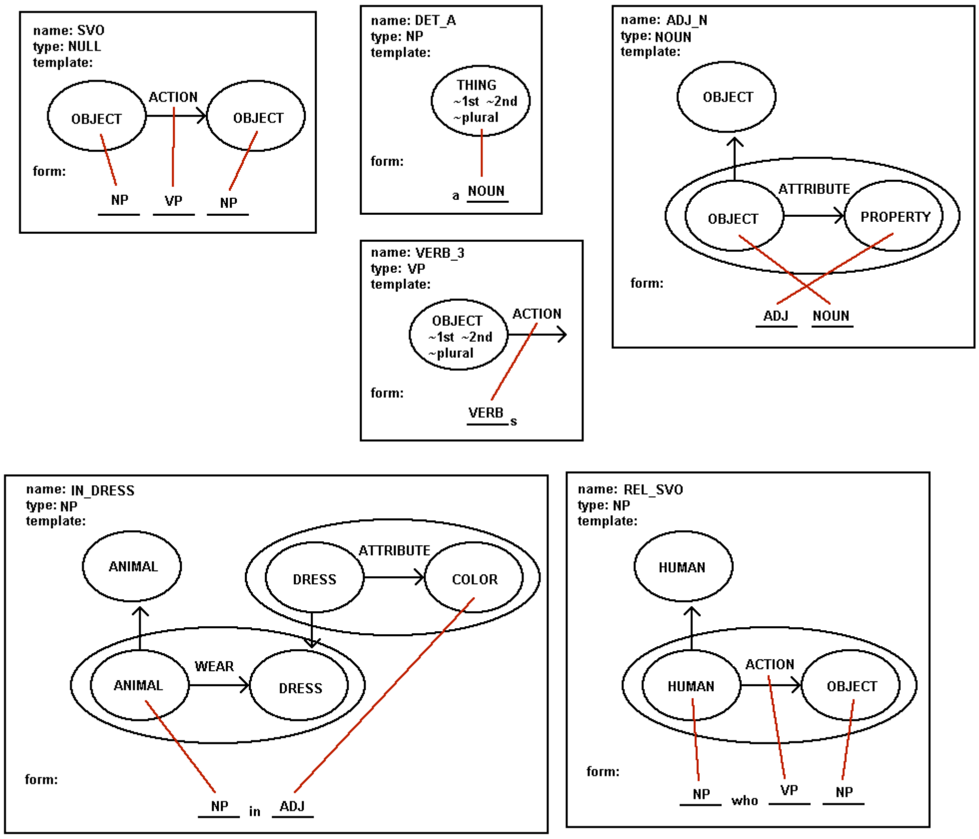

Fig. 6. Abstract constructions used for translation. These constructions are assumed to encode grammatical information.

Figure 6 and Figure 7 provide detailed description for all the constructions used in this example. Some auxiliary information such as activation values, the tense or number is not shown but is assumed to be encoded in the templates (more precisely in the concept attached to the corresponding node or relation) of the constructions. Although activation value is not considered here, it is – as we noted earlier – important in determining the sentence structure – whether it is active or passive. For some constructions, such as SVO or REL_SVO, it is assumed that the activation value for the node corresponding to the agent of an action is higher than that of the patient node and this would lead to produce an active voice. Furthermore, construction VERB_3 is an example of the negation of attributes. Only a single third object is eligible for the conjugation specified in the construction and this grammatical constraint is set by adding negation attributes. Relatively abstract constructions with complex templates and slots in the form are shown in Figure 6 and constructions corresponding to single words are shown in Figure 7. We leave it to the reader to “simulate” the processes of parsing/comprehension and production whereby TCG finds the constructions which convert the SemRep of Figure 2 into the sentence considered here, and vice versa.

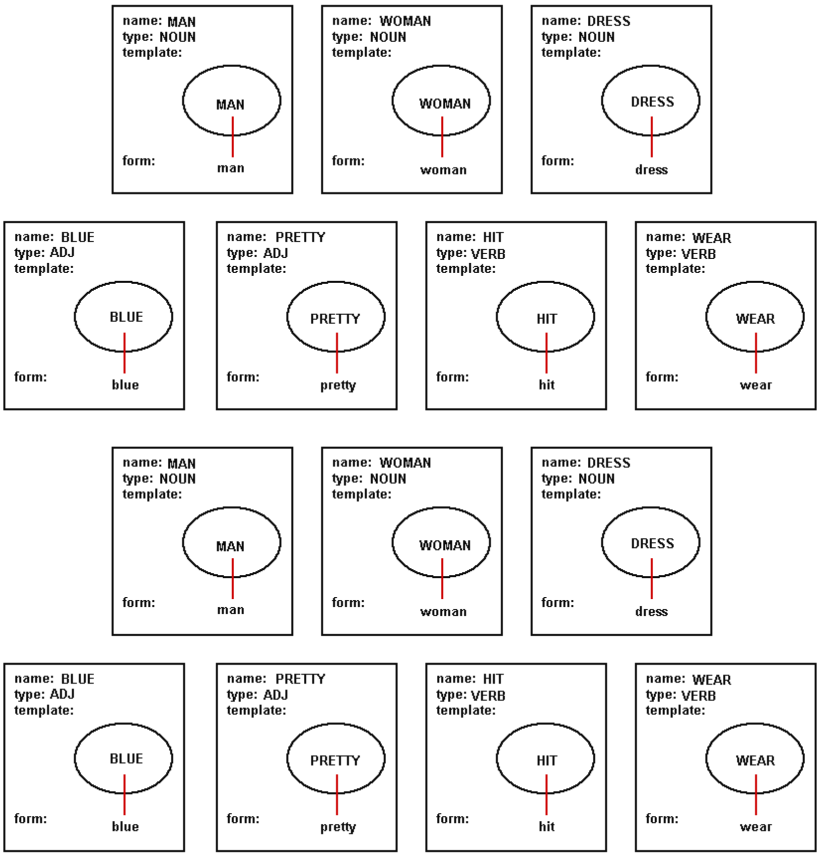

Fig. 7. This figure illustrates the sort of simple construction that corresponds to an element in the lexicon. These constructions are assumed to encode semantic information and can be directly translated into words.

4 Conclusions

4.1 How SemRep Reshapes Construction Grammar

Template Construction Grammar (TCG) shares basic principles with other construction grammar approaches but is explicitly designed to link the semantics of sentences to the representation of visual scenes. However, the use SemRep involves a sufficiently general graphical structure that we are confident of its extensibility to their meanings. SemRep simplifies production and comprehension. Since the task semantics are given as SemRep graphs, the sentence production process is reduced to a general task of matching graphs and the interpreted meaning of a sentence can be directly built by the combination of templates of the activated constructions in the comprehension process.

In addition to template and form pairings, constructions in TCG also encode auxiliary information such as type which specifies the grammatical role that the construction plays. With this information at hand, the language system can build and parse various kinds of grammatical structures appropriate to the task. In any case, the detailed resulting structure is largely dependent on the construction repertoire of the system. The repertoire is maintained in a very dynamic and flexible way, well representing the grammatical constitution and usage pattern that language shows.

Moreover, the concept attached to a node and relation in SemRep graph in TCG formalism exploits the combination of attributes or properties, providing a key comparison mechanism among conceptual entities. During production of sentences, a given graph activates a number of constructions and is compared with a number of constructions for similarity. Only the winner is to be chosen to produce sentences.

On the other hand, in comprehension mode, a textual form is basically what activates constructions by an inverse matching mechanism. In this case, the form, not the template, is what is being compared against the input. When proper constructions are chosen, a new SemRep graph would be built from the templates of the constructions. When multiple constructions are to be combined into a single node or relation, the attributes of the concept of that entity will be added up, getting more specific. In this way, the transformation between different kind of hierarchical structures (back and forth between SemRep and sentence structure) can be executed.

4.2 Another Perspective

The literature on brain mechanisms of vision, and on forms of representation of visual information is, of course vast, and beyond the scope of this article. A subfield of great relevance here is that of vision in embodied agents, with an interest in linking explicit computational analysis of vision in robots to studies of the role of vision in animal and human behavior. Clearly, this field includes our interest in computational models of the control of action and of mirror systems which are involved in both the self's control of action and its vision-based recognition of actions conducted by others. In particular, then, we need to situate our work within the set of studies which unite the study of vision in embodied agents with studies of communication (especially using language) between such agents concerning their visual perceptions (e.g., [22-24]. Another area of concern is discussion of the extent to which construction grammar can be linked to implementations based on neural networks or brain mechanisms (e.g., [25, 26]). However, in this paper, we restrict our discussion to one paper, [27], from the group of Luc Steels, a group which has not only been a leader in linking the study of vision in embodied agents with studies of communication, but has done so within the framework of simulated evolution (though not linked to neurobiology), and has developed its own version of construction grammar, Fluid Construction Grammar (FCG).

Steels and Loetzsch [27] use interactions between robots to study the effect of perspective alignment on the emergence of spatial language. Although the authors state that their "experiments rest on the Fluid Construction Grammar framework [21], which is highly complex software for language processing", there is no syntax in the language studied in their paper – rather, visual scenes are described simply by a list of words which are associated with one or more categories applicable to the observed scene. We postpone a comparison (and, perhaps integration) of TCG and FCG for another occasion, and instead focus on scene representation in [27] and then compare it with SemRep to help clarify directions for future work.

[27] employs an actual vision system to generate scene descriptions from visual input provided by cameras mounted on 2 or more AIBO robots. In a typical episode, two robots are placed in a cluttered room and move about till each has both the other robot and a ball in their visual field; they then stay still while a human uses a stick to move the ball from one position to the other. Each robot generates a description of the ball's trajectory using Cartesian coordinates for the ground plane, with the robot at the origin and its direction of gaze determining the vertical axis. The descriptors given are

(1) x of start point, distance to start point, x of end point, y of end point, distance to end point, angle to end point, angle of movement, length of trajectory, change in x, change in y, change in angle, and change in distance.

The key property of language addressed in [27] is that of *perspective alignment* – different observers may describe the same scene in different terms – does "on the left", for example, mean "on the speaker's left" or "on the hearer's left"? To address this challenge, each robot is programmed to use its vision to judge the position and orientation of the other robot and then estimate the above coordinates (1) as seen from the other robot's viewpoint. This *perspective transformation* is a simple translation and rotation in Euclidean space, but the result is an estimate because the robot's assessment of the relevant coordinates may contain errors and these are unlikely to correlate with errors of the other robot.

Neither words nor categories for describing the scene are provided in advance. Rather, simple discriminant trees are used to create categories: every feature in (1) has a discrimination tree which divides the range of possible values into equally sized regions, and every region carves out a single category. Letter strings can be randomly generated to provide "words", and weighted, many-to-many links between words and categories can be stored in a bidirectional associative memory [28]. However, from this random initial state, interactions between 2 or more robots allow them to end up with a set of categories, and a set of words associated with those categories, that allow any 2 robots to communicate effectively about a scene, adopting either their own perspective or that of the other robot. As noted, each “utterance” consists of a small set of words; these activate certain categories. A robot will strengthen its current “knowledge” if it can match the word string it “hears” to the scene it “sees” or to its estimated perspective for the other. If neither match is possible, it will change its categories and/or vocabulary and/or bidirectional association between words and concepts to better match one perspective with the utterance.

More specifically, learning extends over thousands of episodes. After a successful exchange, the score of the lexical entries that were used for production or parsing is increased by 0.05. At the same time, the scores of competing lexical entries with the same form but different meanings are decreased by 0.05 (lateral inhibition). In case of a failure, the score of the involved items is decreased by 0.05. This adjustment acts as a *reinforcement learning* mechanism and also as *priming* mechanism so that agents gradually align their lexicons in consecutive games. Similar mechanisms apply to the updating – and eventual alignment – of categories in each robot on the basis of success or failure in each exchange.

With this, we use our understanding of [27] to sharpen our understanding of SemRep and to pose challenges for future research:

Rather than use a very limited type of description –how the same object, the ball, moves in each episode – we are concerned with a flexible description of an episode, or small number of contiguous episodes, that labels the visual field with concepts related to agents, objects and actions and their attributes, and links them in hierarchical ways. In other words, where [27] focuses on a single intransitive movement (the ball rolls), we have a special concern with transitive actions, based on evaluating the movement of an agent with respect to an object or other agent.

We have not implemented a vision front-end, but note that in fact the language-related work in [27] does not make essential use of the vision front-end, since the “real processing” starts with the Cartesian coordinates provided in (1) both from the robot’s own perspective and as estimated for the other robot’s perspective. In terms of the VISIONS system [4], this would correspond to the converged state of the intermediate database, but rather than giving coordinates of a single trajectory in the ground plane, an extension of VISIONS would label shapes and edges and their relative position and motion in the three-dimensional visual field of the observer. Just as [27] uses this description as the basis for extracting a small set of categories, so we would use the intermediate database as the basis for constructing a SemRep, while noting that the choice of SemRep may depend on attentional factors and task relevance [6, 29], including the state of discourse.

Concepts and words are emergent in [27] through attempts to share descriptions of observed scenes. SemRep uses hand-crafted concepts, words and constructions.

Perspective-taking is almost obligatory in [27] – in all but one experiment (see below) each robot must compute the description (1) as seen by the other robot. In SemRep, we do not use any such global transformations, but rather rely on a set of appropriate “subscene schemas”, so that a portion of the same SemRep could be described by “the man to the left of the woman” or, if we take into account the orientation of the woman’s body, “the man in front of the woman.” We note the further challenge of deciding when two SemReps could apply to the same scene as viewed from different perspectives (perhaps with different foci of attention) or, more subtly, could describe two different time slices” of a spatially extended scene.

“Cognitive effort” is defined in [27] as the average number of perspective transformations that the hearer has to perform. Their Figure 12 reports an experiment which shows (perhaps unsurprisingly) a marked reduction in cognitive effort when perspective is marked, i.e., when one of the categories that must be expressed is whether the trajectory descriptors in (1) are based on the perspective of the “speaker” or the “hearer”. In this experiment, separate words emerged for perspective in addition to words where perspective is part of the lexicalization of the predicate. Steels and Loetzsch [27] comment that “This is similar to natural language where in *the ball to my left*, *my* is a general indicator of perspective, whereas in [...] *come* and *go*, perspective is integrated in the individual word” and assert that “this experiment explains why perspective marking occurs in human languages and why sometimes we find specific words for it.” However, the experiment does not explain this directly, since the choice of perspective is added by the authors as an explicit category – thus making it likely that words will emerge to express or incorporate this category. However, an important evolutionary point is made: if the perspective category or word is available (whether through biological or cultural evolution) then processing is more efficient, thus giving creatures with access to such items a selective advantage. When we turn from robot routines to human development, the question is how the child comes to recognize its similarity and difference from others so that terms like “my left hand” versus “your left hand” become understood, and then how such spatial terms extend from the body to peripersonal space and then to space generally. It is not surprising that – just as in the language games described here – different languages will describe this extension in different ways.

We close by a (perhaps surprising) link from the present discussion back to our earlier concern with models of the mirror system. Figure 10 of [27] summarizes an experiment in which the agents perceive the scene through their own camera but they “do not take perspective into account.” In this case, the agents do not manage to agree on a shared set of spatial terms. Steels and Loetzsch concludes that this proves that “grounded spatial language without perspective does not lead to the bootstrapping of a successful communication system.” However, this does not take account of the extent to which the results depend on what is built into the system. Other approaches are possible. Suppose the room had several distinctive landmarks. Then instead of locating the ball in one of the two prespecified Cartesian coordinate systems, one could locate the ball in terms of descriptions like “It started close to landmark-3 and moved halfway to landmark-7.” (In neural terms, such a description might build on the activity of place cells in the hippocampus [30].) Here no perspective transformation is involved. The latter approach is more like that taken in the MNS models [9, 10]. Instead of describing the movement of the hand in, e.g., retinal coordinates, we there described it in object-centered coordinates, thus eliminating the issue of perspective-taking. Of course, this does not guarantee that our assumption is justified. However, one argument in favor of (but not proving) the assumption is that the need for visual feedback for dexterity would provide selection pressure for a system that could translate retinal input into such an object-centered (or affordance-based) view of the hand.

5 References

[1] Arbib, M.A.: Rana computatrix to human language: towards a computational neuroethology of language evolution. Philos Transact A Math Phys Eng Sci. 361, 2345-2379 (2003)

[2] Arbib, M.A., and Liaw, J.-S.: Sensorimotor Transformations in the Worlds of Frogs and Robots. Artificial Intelligence. 72, 53-79 (1995)

[3] Arbib, M.A.: Perceptual structures and distributed motor control. In: Brooks, V.B., (ed.) Handbook of Physiology — The Nervous System II. Motor Control 1449-1480, American Physiological Society, Bethesda, MD (1981)

[4] Draper, B.A., Collins, R.T., Brolio, J., Hanson, A.R., and Riseman, E.M.: The schema system. International Journal of Computer Vision. 2, 209-250 (1989)

[5] Erman, L.D., Hayes-Roth, F., Lesser, V.R., and Reddy, D.R.: The HEARSAY-II speech understanding system: Integrating knowledge to resolve uncertainty. Computing Surveys. 12, 213-253 (1980)

[6] Itti, L., and Arbib, M.A.: Attention and the minimal subscene. In: Arbib, M.A., (ed.) Action to language via the mirror neuron system 289-346, Cambridge University Press, Cambridge (2006)

[7] Fagg, A.H., and Arbib, M.A.: Modeling parietal-premotor interactions in primate control of grasping. Neural Netw. 11, 1277-1303 (1998)

[8] Gibson, J.J.: The ecological approach to visual perception. Houghton Mifflin, Boston (1979)

[9] Bonaiuto, J., Rosta, E., and Arbib, M.: Extending the mirror neuron system model, I : Audible actions and invisible grasps. Biol Cybern. 96, 9-38 (2007)

[10] Oztop, E., and Arbib, M.A.: Schema design and implementation of the grasp-related mirror neuron system. Biol Cybern. 87, 116-140 (2002)

[11] Gershkoff-Stowe, L., and Goldin-Meadow, S.: Is there a natural order for expressing semantic relations? Cognitive Psychology. 45, 375-412 (2002)

[12] Arbib, M.A., and Rizzolatti, G.: Neural expectations: a possible evolutionary path from manual skills to language. Communication and Cognition. 29, 393-424 (1997)

[13] Arbib, M.A.: From Monkey-like Action Recognition to Human Language: An Evolutionary Framework for Neurolinguistics (with commentaries and author's response). Behavioral and Brain Sciences. 28, 105-167 (2005)

[14] Arbib, M.A.: Broca’s Area in System Perspective: Language in the Context of Action-Oriented Perception. In: Grodzinsky, Y., and Amunts, K., (eds.) Broca’s Region 153-168, Oxford University Press, Oxford (2006)

[15] Saussure, F.: Cours de linguistique générale (ed. by C. Bally and A. Sechehaye, with the collaboration of A. Riedlinger). Payot Lausanne and Paris (1916) (English translation by W. Baskin, Course in General Linguistics, Fontana/Collins, Glasgow 1977)

[16] Gallese, V., and Goldman, A.: Mirror neurons and the simulation theory of mind-reading. Trends Cognit. Sci. 2, 493-501 (1998)

[17] Chomsky, N.: Lectures on Government and Binding. Foris, Dordrecht (1981)

[18] Croft, W., and Cruse, D.A.: Cognitive Linguistics. Cambridge University Press, Cambridge (2005)

[19] Goldberg, A.E.: Constructions: A new theoretical approach to language. Trends in Cognitive Science. 7, 219-224 (2003)

[20] Fillmore, C.J., Kay, P., and O'Connor, M.K.: Regularity and idiomaticity in grammatical constructions: the case of let alone. Language & Cognitive Processes. 64, 501-538 (1988)

[21] De Beule, J., and Steels, L.: Hierarchy in Fluid Construction Grammar. In: Furbach, U., (ed.) KI-2005. Lecture Notes in AI 3698, 1-15, Springer-Verlag, Berlin (2005)

[22] Roy, D.: Semiotic schemas: A framework for grounding language in action and perception. Artificial Intelligence. 167, 170-205 (2005)

[23] Steels, L.: Evolving grounded communication for robots. Trends in Cognitive Sciences. 7, 308-312 (2003)

[24] Cangelosi, A., and Riga, T.: An Embodied Model for Sensorimotor Grounding and Grounding Transfer: Experiments With Epigenetic Robots. Cognitive Science. 30, 673-689 (2006)

[25] Dominey, P.F., and Hoen, M.: Structure mapping and semantic integration in a construction-based neurolinguistic model of sentence processing. Cortex. 42, 476-479 (2006)

[26] Feldman, J., and Narayanan, S.: Embodied meaning in a neural theory of language. Brain Lang. 89, 385-392 (2004)

[27] Steels, L., and Loetzsch, M.: Perspective Alignment in Spatial Language. In: Coventry, K.R., Tenbrink, T., and Bateman, J.A., (eds.) Spatial Language and Dialogue, Oxford University Press, Oxford (2007)

[28] Kosko, B.: Bidirectional associative memories. IEEE Transactions on Systems, Man and Cybernetics. 18, 49-60 (1988)

[29] Navalpakkam, V., and Itti, I.: Modeling the influence of task on attention. Vision Research. 45, 205-231 (2005)

[30] Guazzelli, A., Corbacho, F.J., Bota, M., and Arbib, M.A.: Affordances, Motivation, and the World Graph Theory. Adaptive Behavior. 6, 435-471 (1998)