available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/brainres

BRAIN RESEARCH

Research Report

Describing visual scenes: Towards a neurolinguistics based on construction grammar

Michael A. Arbib[a,b,c,*], *JinYong Lee*[a]

[a]*Computer Science, University of Southern California, Los Angeles, CA 90089-2520, USA*
[b]*Neuroscience, University of Southern California, Los Angeles, CA 90089-2520, USA*
[c]*USC Brain Project, University of Southern California, Los Angeles, CA 90089-2520, USA*

ARTICLE INFO

Article history:
Accepted 21 April 2008

Keywords:
Action
Action recognition
Brain mechanism
Competition and cooperation
Construction grammar
Dynamic visual scene
Language perception
Language production
Mirror neuron
Visual perception
Scene description
Schema theory
SemRep
Vision

ABSTRACT

The present paper is part of a larger effort to locate the production and perception of language within the broader context of brain mechanisms for action and perception more generally. Here we model function in terms of the competition and cooperation of schemas. We use the task of describing visual scenes to explore the suitability of Construction Grammar as an appropriate framework for a schema-based linguistics. We recall the early VISIONS model of schema-based computer analysis of static visual scenes and then introduce SemRep as a graphical representation of dynamic visual scenes designed to support the generation of varied descriptions of episodes. We report preliminary results on implementing the production of sentences using Template Construction Grammar (TCG), a new form of construction grammar distinguished by its use of SemRep to express semantics. We summarize data on neural correlates relevant to future work on TCG within the context of neurolinguistics, and show how the relation between SemRep and TCG can serve as the basis for modeling language comprehension.

1. Introduction

To approach the general issue of how the brain can go back and forth between semantic representations and the utterances of some language, we focus on how one may go from a visual scene to a description of that scene. We briefly outline how the brain may be modeled in terms of the competition and cooperation of functional entities called schemas and then present an argument concerning the linkage between a mirror system for words, a mirror system for actions, and neural mechanisms supporting the processing of perceptual and motor schemas. We then complete our tour of necessary background by recalling key features of the VISIONS model of schema-based computer analysis of static scenes. The remainder of the paper is not about neurolinguistics *per se*, but rather it offers hypotheses on how visual and linguistic structures may be represented in such a way that each may be mapped into the other through schema interactions. It will be

* *Corresponding author*. USC Brain Project, University of Southern California, Los Angeles, CA 90089-2520, USA.
E-mail addresses: arbib@usc.edu (M.A. Arbib), jinyongl@usc.edu (J. Lee)

0006-8993/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.brainres.2008.04.075

the task of future papers to offer explicit hypotheses about the localization of these processes in the human brain.

Section 2 explains how the processes developed in VISIONS for the competition and cooperation of schemas in the analysis of static visual scenes may be extended to the analysis of episodes in the interpretation of ongoing visual experience. This extension of VISIONS provides the input for SemReps, graphical representations of dynamic visual scenes which can support the generation of varied descriptions of episodes. Section 3 then 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 which express the information encoded in SemReps. Section 4 then works through a detailed example of how TCG can operate to convert a SemRep into a sentence. Finally, Section 5 compares TCG with other models of CG, summarizes data on neural correlates of vision and language relevant to future work on TCG within the context of neurolinguistics, and shows how the relation between SemRep and TCG can serve as the basis for modeling language comprehension.

1.1. *Schemas which compete and cooperate*

In the present paper, we approach brain mechanisms of vision and language through the analysis of "schemas" as the "distributed programs" of the brain, at a level above, but reducible to, the functioning of neural networks. For example, we have *perceptual schemas* recognizing apples and doors, and *motor schemas* for peeling apples and opening doors. Each of these schemas constitutes a psychological reality and we can combine schemas into coordinated control programs and schema assemblages (Arbib, 1981; Arbib et al., 1998) to define complex courses of action. It may help to note the distinction between basic and neural schema theory:

Basic Schema Theory works at a functional level which associates schemas with specific perceptual, motor, and cognitive abilities and then stresses how our mental life results from the dynamic interaction – the competition and cooperation – of many schema instances. It refines and extends an overly phenomenological account of the "mental level". There are echoes here of the use of the term *schema* in neurology (Head and Holmes, 1911), psychology of memory (Bartlett, 1932), kinesiology (Schmidt, 1975), and genetic epistemology (Piaget, 1971).

Neural Schema Theory provides the "downward" extension of basic schema theory as we seek to understand how schemas and their interactions may indeed be played out over neural circuitry – a basic move from psychology and cognitive science as classically conceived (viewing the mind "from the outside") to cognitive neuroscience. However, the linkage of schemas to brain regions may serve not only as the framework for modeling neural circuitry but may also ground models "at the schema level" which can be tested against lesion data and brain imaging (Arbib et al., 1998), a level which seems particularly relevant to neurolinguistic studies.

The notion of *coordinated control program* (Arbib, 1981) shows how to build up complex skills from available perceptual and motor schemas (where the perceptual schemas may register the external environment, or the degree of achievement of goals and subgoals), and includes specification of both how data are to be transferred between schemas and how schemas are to be activated and deactivated. In particular, perceptual schemas may estimate the value of parameters relevant to the way in which motor schemas control ongoing action. Lyons and Arbib (1989) provided a complete formalism for the combination of schemas into coordinated control programs, and the resultant RS (Robot Schema) language has been used to develop a range of programs for embodied, perceptually-guided behavior of robots. However, in this paper we focus on the way in which schemas may be associated with a visual scene to yield a Semantic Representation (SemRep) that can be used as the basis for generating a verbal description of the scene. It remains a challenge for future research to integrate the RS formalism with the tools for sentence construction of Template Construction Grammar (TCG) presented in Section 3.

In this paper, we focus on the role of vision in segmenting a scene and labeling the regions, or detecting characteristic patterns of motion in a videoclip to provide a semantic representation which can challenge our research on brain mechanisms of language. However, this is just one facet of a broader approach to vision which is concerned with its relevance to the ongoing behavior of an embodied agent — be it frog, rat, monkey, human or robot (Arbib and Liaw, 1995; Arbib, 2003). 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.

The notion of *schema assemblage* relates to the observation (Arbib and Didday, 1971) that perception of a scene may be modeled as invoking instances of perceptual schemas for certain aspects of the scene, but not simply as discrete labels — each schema instance has parameters related to size, location and other features of the represented element of the scene. Here, a perceptual schema is the knowledge in long-term memory of, e.g., how to recognize a chair when one sees one, whereas our working memory of a scene will contain a separate schema instance for /chair/ for each region that application of that knowledge assesses as being, with some level of confidence, a chair. The emerging pattern of schema instances which comes to represent the scene (and thus give it its current meaning to the observer) may result from extensive processes of competition and cooperation in the schema network which invoke schemas beyond those initially associated with the scene. *Cooperation* occurs in, for example, the mutual increase of the confidence level of schema instances for different regions of the image if each provides a plausible context for the other – the schema for "foliage" gets a boost for interpreting the region just above a region already interpreted as a tree trunk, and vice versa. *Competition* occurs when schemas compete to interpret a particular region of a scene — "Is that a bird, is it a plane? No, it's Superman." Thus, a schema instance may initially become more active if it is consistent with more features of a region which it is competing to interpret. Cooperation then yields a pattern of "strengthened

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

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 representation of the environment in working memory.

1.2. *From mirror neurons to neurolinguistics*

The general conceptual model of Fig. 1 (Arbib, 2006) links brain mechanisms for the perception and production of language to more general mechanisms for scene perception and praxic action, and – while much broader in scope – is consistent with the analysis of the roles of the dorsal and ventral streams in speech processing posited by Hickok and Poeppel (2004, 2007). The top half of the figure summarizes the general notion of the Mirror System Hypothesis (Arbib, 2005; Rizzolatti and Arbib, 1998), namely that the dorsal stream in humans contains mirror systems for both praxic actions and for words considered as articulatory-actions (which may include manual and facial as well as vocal production). However, the model posits that the relation between the two systems is not one of direct connectivity in which activating the mirror neuron encoding of an action elicits the mirror neuron representation of a corresponding verb. Indeed, relatively few words are verbs for actions within the speaker's repertoire and so relatively few can have their neural representation linked to mirror neurons for actions.

The relation with the upper and lower portions of Fig. 1 is inspired by the FARS model (Fagg and Arbib, 1998) of how prefrontal influences (e.g., task knowledge or working memory) determined via the ventral stream may affect which motor schema for an action compatible with affordances selected by parietal cortex, as determined by the dorsal stream, will be brought above threshold for execution in premotor cortex. We posit that the dorsal mirror system for words-as-articulatory-actions is linked to a system in the ventral stream and prefrontal cortex that supports the action-oriented perception which builds a schema-based representation of the external world in the service of planning the ongoing behavior of the organism.

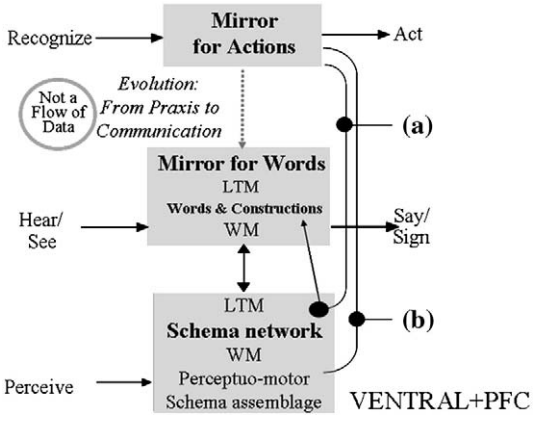

Fig. 1 – (from Arbib, 2006). Words link to schemas, not directly to the dorsal path for actions. Consider, for example, the differential linkage of nouns and verbs to perception of the objects and actions, respectively, that they denote.

The notion, then, is that the ventral pathway selects which actions to deploy and (via pathway (a)) activates the parameterized motor schemas to execute those actions. According to the Mirror System Hypothesis, in humans evolution has "lifted" this set of processes so that a schema assemblage (which in this case may be far more abstract than linking the current scene with a plan of action) can ground the planning of utterances which can in turn deploy and (via pathway (b)) activate the parameterized motor schemas to articulate the words of that utterance.

Our concern in the present paper is with two questions concerning the lowermost box of Fig. 1:

1a) How is the schema assemblage that represents the current scene or (more generally) the recent history of the current scene to be represented in a form which grounds verbal description of the scene?
1b) How is this representation to be converted into one or more sentences?

We will expand here upon answers provided earlier (Arbib and Lee, 2007):

2a) SemRep (Section 2) will provide the semantic structure for scene description.
2b) Template Construction Grammar (Section 3) will provide the mechanisms for generating descriptions corresponding to a given SemRep.

1.3. *The VISIONS system*

An early example of schema-based interpretation for visual scene analysis is the VISIONS system (Draper et al., 1989) which deploys a set of perceptual schemas to label objects in a static visual scene. In VISIONS, the general nature of the scene (e.g., an outdoor scene with houses, trees, lawn, etc.) is prespecified, and only those schemas are deployed which are relevant to recognizing this kind of scene. Nonetheless, the challenge of recognizing what object is located where in the scene remains daunting, and distressingly little work has been done on general mechanisms of visual scene analysis since the VISIONS effort.

When a new image is presented to the VISIONS system for processing, low-level processes akin to those at early stages of the mammalian visual cortex build 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 not static, but may change as the process of interpretation proceeds. This is because it 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). More generally (though this was outside the scope of the VISIONS effort), the analysis of the image will depend on how attention is focused on different parts of the image, and with different attention to

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

detail; and this in turn will depend on the current goals or task requirements of the observer (Itti and Arbib, 2006).

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 WM (working memory — Draper et al., 1989, refer to this as STM, short-term 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.

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. Interpretation of a novel scene starts with the data-driven instantiation of several schemas (e.g., a certain range of color and texture in part of the image might cue an instance of the foliage schema). 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 WM has an associated confidence level which changes on the basis of interactions with other units in WM. The WM 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, as activated schema instances 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 (parts of) the scene is obtained.

2. SemRep: A semantic representation for dynamic visual scenes

The range of speech acts is immense — we can request, cajole, lie, inform and ask questions, to name just a few. Here we focus on one type of speech act that must have held great importance in the evolution of language — the ability to describe to another those aspects of the environment that have caught one's attention. Such a description provides a compact representation of the scene that is readily communicable as a string of words, and thus cannot do justice to all the nuances of a scene, though one or two may be singled out for explicit mention. Our first step, then, is to find an economical semantic representation of a visual scene that is directly related to the structure of schema instantiations returned by neural processes akin to those of the VISIONS system, and yet can serve as the basis for generating a sentence according to some grammar. Since we move beyond VISIONS to dynamic scenes, we introduce timeline considerations. We have two 2D images at each time or, equivalently, a $2^1/2$ D sketch (Marr, 1982) from which the 3D (and temporal) distribution of objects, actions and attributes is to be inferred. Whether for vision or language, the various representations are evanescent, and the brain must employ working memory of certain relevant prior activity, and a "working precognition" that holds expectations, goals and other activity relevant to possible courses of action. In neither case should one expect the necessary activity to be gathered in a single neural structure.

In VISIONS, the intermediate data base is dynamic even for static visual input. It provides a current estimate of relevant information about edges, regions, shapes, colors, textures, etc. The input is fixed and interpretation continues until the image is interpreted. However, shifting attention or task demands can change the interpretation by attending to details or regions that had not yet been fully analyzed. The top-level of VISIONS provides a set of schema instances each linked to a certain region of the image, each provided with certain cross-linkages (e.g., a roof is part of a house) as well as certain parameters. Let us now extend this to dynamic scenes:

2.1. *Intermediate database*

At any time t, there will be a state $I(t)$ of the intermediate database, which assigns to each point in the visual field a set of features, and also demarcates regions and boundaries, providing descriptors such as orientation and contrast along each boundary, and information about color, depth, shape and motion for each region. However, $I(t)$ does not interpret the regions. It changes in response both to changes in retinal activity (depending on changes of light from the external world as well as changes in eye position, vergence, accommodation, etc.) and to top-down influences from schema processing.

2.2. *Visual working memory*

At each time, visual working memory $SI(t)$ will contain a set of schema instances, with each associated with a region (or union of regions) delimited in $I(t)$, with each schema instance having an associated confidence level, and an associated set of parameter values which are specific to that schema but which may be shared with sets of related schemas. Change in $SI(t)$ is relatively conservative. If the region to which a schema instance in $I(t)$ is linked remains (even if changed somewhat) in $I(t)$ then that schema instance will probably remain in $SI(t')$ for t' somewhat greater than t, but with a change in confidence level based on the dynamics of competition and cooperation. However, discontinuities arise, as shifting attention may render certain schema instances as no longer of interest while other new ones may be invoked.

2.3. *Minimal or anchored subscene*

The notion introduced by Itti and Arbib (2006) is that once an object or action has become of sufficient interest, it will act as an "anchor" for directing attention in such a way as to find other aspects of the scene which are related to that anchor. Thus, among the schema instances which are most active at a given time, there will be a certain number which cohere to define a "scene" or "episode". We thus divide time into alternating periods (t_n, s_n; w_n) such that an anchor is defined at time t_n, the anchored subscene is developed through schema instance competition and cooperation by s_n, and

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

remains the dominant coalition of schema instances until time w_n. Note, then, that, the set of schema instances above any given threshold value at time t, may constitute zero, one or more anchored subscenes, with several "spare" schema instances that have not been linked to other instances to form a subscene but may yet (but need not) provide anchors for later formation of novel subscenes.

Given this, we introduce SemRep as an encapsulation of what is in visual working memory which is organized around the notion of anchored subscenes. We will define SemRep below as 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 abstraction from the schema assemblages generated by the VISIONS system but with the crucial addition of actions and events extended in time. A cautionary note: the analysis of single images in VISIONS is only the "opening wedge" for dynamic scene analysis (Itti and Arbib, 2006). Thus, if we see a single picture of a man with a woman's hand flat against his cheek, and we have no cues from his facial expression, we could not tell whether the scene is "woman slapping man" or "woman stroking man's cheek" unless we could see the temporal sequence of which this is part. In this regard, note that the MNS model of action recognition (Bonaiuto et al., 2007; Oztop and Arbib, 2002) is based on recognizing the temporal trajectory of a hand relative to an object. This model can serve as the basis for other models of action recognition, as indicated by the generalized features of our model summarized as follows:

2.4. *Action recognition schema*

An *action recognition schema* takes as inputs visual data concerning two trackable regions of the image, $R_1(t)$ and $R_2(t)$, and invokes perceptual schemas to classify the two regions as B_1 and B_2, and use this classification to track certain key features of B_1 and B_2 as they appear in $R_1(t)$ and $R_2(t)$. The action recognition schema then uses relative trajectories of certain of these features as the basis for classifying the action which is taking place.

For example, in the MNS model, one region must be recognized as a hand and the other as an object with the corresponding features being based on wrist and finger positions for the hand and affordances (position of surfaces to which an action may be applied, in this case) for the object. The system then takes the trajectory of hand features as expressed in an object-centered framework, and returns a confidence level for different actions such as precision pinch and power grasp, with (in general) the confidence level for a single action rising well above the others as the action moves towards completion.

Thus, an action involves a dynamic region of visual space which contains the regions for (for example) the agent and patient of the action, and the space in which they move together. In the version of SemRep presented here, we represent such a situation as an edge, associated with an action, which links nodes for the agent and patient of the action. Future work will explore alternatives in which a node for the action is linked to the bounding region of the action, with an agent-link and a patient-link to nodes which are linked to the regions of the scene corresponding to the agent and action, respectively.

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.

In VISIONS, the schema instance level may invoke the intermediate database to in turn invoke further processing to return answers needed to assist the competition and cooperation between schema instances, so we must understand that SemRep is not an isolated graphical representation but is instead linked to the schema instance level but gives only a partial view of it. Each node is linked to a region of the image either via the schema instance for that region (e.g., for an object) or to parameters describing that region (as when an attribute is linked to a node corresponding to that region) or to the linkage between regions related to agents or objects (as in the case of actions or spatial relations) which may encompass a somewhat larger region.

We noted earlier that a schema may be associated with a number of parameters. Some of these parameters may be relevant to possible interactions with an object which the schema represents, for example, and yet not be available to verbal expression or pantomime (Goodale and Milner, 1992). But even those parameters which could be available may not be available, and will only become available if they are needed for cognitive processing, such as the planning (rather than the execution) of action, for problem-solving, or for verbal description. However, just as VISIONS shows how the demands of schema instantiation can generate new processing requests back down to the intermediate data base, so we postulate that SemRep need explicitly represent very few parameters, and can direct requests to Visual Working Memory when the information is needed for cognitive processes. In other words, while some parameters may be explicitly represented at this level, others are simply "available" at this level in that enquiries addressed to SemRep may be passed to the schema instance level. Such enquiries may, if necessary, be routed via spatial covering back to lower levels which contain more precise information on, e.g., shape or the distribution of color or texture. Each parameter which does get made explicit at the SemRep level is considered an attribute and given its own node to be linked to the node for the parameterized schema.

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

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

necessary. On the other hand, a Gricean convention (Grice, 1969) 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. Here one may need to include what is not in the image to block standard inferences where they are inappropriate. This is akin the notion of inheritance in semantic networks.

Building a SemRep may require multiple fixations of the same region, both to accumulate descriptors for the region and to extract "relations" between this and other regions. If either fixation shown in Fig. 2(1) were made first, the only information associated with it might be "house". But once we have seen both, we may want to distinguish them. This may require an alternation of fixations to settle on salient characteristics such as color, and thus distinguish them by the labels "house/red" and "house/blue". In addition, or instead, the alternating fixations might establish a spatial relation between them as in Fig. 2(2). But note that once we have all this information, we can use it in very different ways. For example, we might ignore the color difference, and use the spatial relation to speak of "the house on the left" rather than "the red house". Alternatively, we might want to describe the subscene in Fig. 2(2) by saying "There's a red house to the left of a blue one" or "There's a blue house to the right of the red one".

Moreover, in forming a description of part of the scene, the whole SemRep may affect the expression of the part. Based on the SemRep in Fig. 2(3), we might say "There's a kid playing in front of the house". However, if our analysis of the scene had already yielded the SemRep in Fig. 2(4), then in describing the subscene in Fig. 2(3) we are obliged to specify which house is involved, and (perhaps after refining the analysis of each house, in the fashion described in discussing Fig 2(1)) say more specifically "There's a kid playing in front of the red house."

In response to hearing "There's a kid playing in front of the house." one might ask "Who's the kid?" This might involve retrieving the node [1] in Fig. 2(5) and then *narrowing attention* to [2] – localizing the kid – and then to [3] – looking at his face to recognize that it is "John". (Note the important distinction between "shifting attention" and "narrowing attention".) The language system could then answer the question in various ways:

> "John.", "The kid's John." "John is the kid". "It's John." And so on.

At this stage, node [3] might be dropped from the SemRep, having served its role in face identification, but node [2] might be maintained as part of the current SemRep and inherit the label "John". This might trigger further analysis of [1], shifting attention to [4] in Fig. 2(6), perhaps on the basis of John's outstretched arm, to recognize the ball and to then link [2] and [4] via action recognition (perhaps invoking the mirror system), resulting in a graph that could be verbalized as "John is hitting the ball."

In verbalizing a given part of the SemRep, we may choose how much we refine a node into subgraphs that it dominates, e.g., to say "A kid is playing in front of a house", "John is hitting the ball in front of the red house", "John is playing in front of the house on the left", "John is playing in front of one of the houses", etc. Of course, other fixations could yield extensions of the SemRep that could be expressed with other sentences, such as "John is near the tree" (perhaps in answer to the question "Where's John?" — an alternative answer, of course, being "In front of the red house".) Again, if we followed the arc of the ball and saw Bill (not shown in the figures) we might extend the sentence "John is hitting the ball" to "John is hitting the ball to Bill" or "John is hitting Bill the ball."

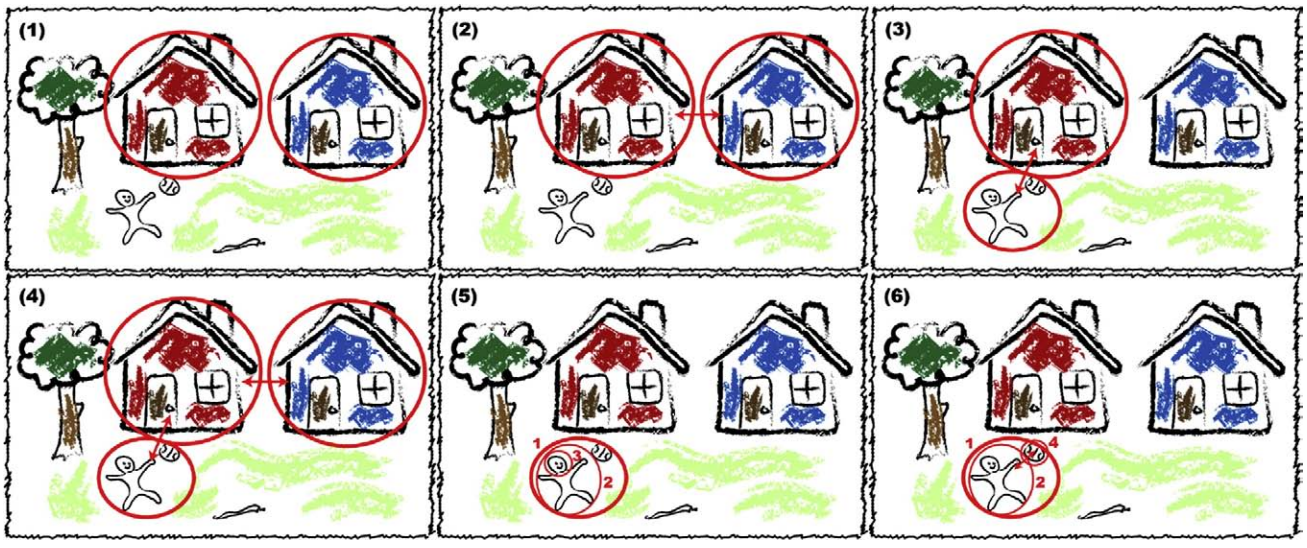

Fig. 2 – Different patterns of attention to a visual scene will highlight different regions of interest and relations between them, defining a graph structure called SemRep. We stress that a given visual scene may support a large number of SemRep graphs, and that a single SemRep may support many different sentences to describe those aspects of the scene which it represents.

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

One more observation. The processing of schemas by VISIONS includes the setting and updating of "confidence levels", indicating the weight of evidence for a specific schema instance to interpret a given region, based on the state of competition and cooperation within the network. Similarly, in SemRep, each node is assigned a value representing "discourse importance" — what the speaker wishes to emphasize for the hearer. For instance, even if the vision system had specified BOY first and this led to the zooming in on the face, FACE might be ranked higher in SemRep than BOY if the former is currently of greater interest. Again, 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. The use of these weights is already included in the design of our Template Construction Grammar, but we will not discuss them further in this paper.

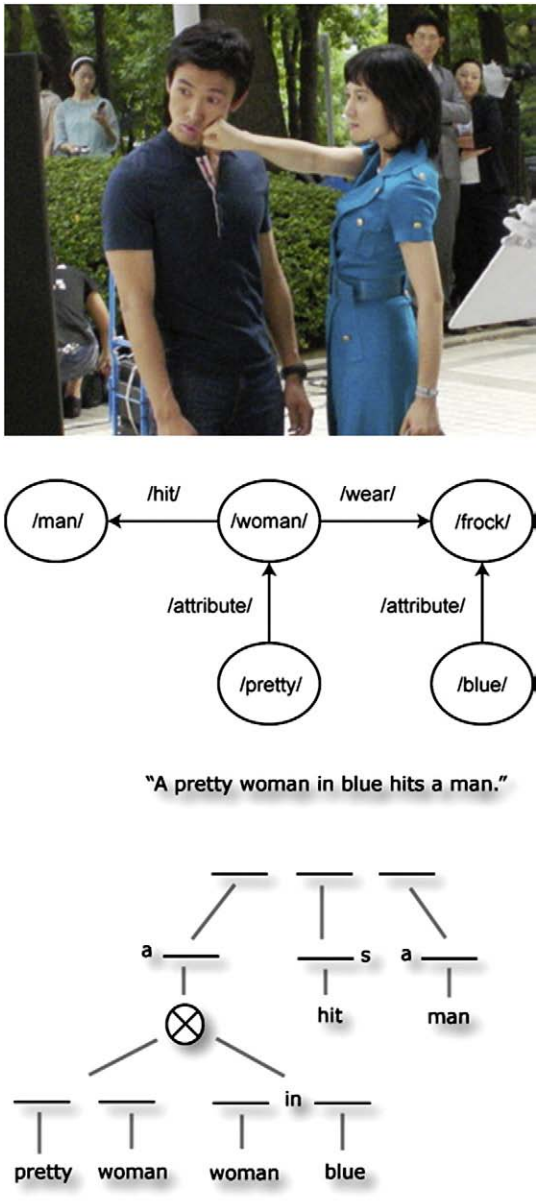

Fig. 3 – Top: A picture of a woman hitting a man (original image from "Invisible Man Jangsu Choi", Korean Broadcasting System). Middle: A SemRep graph that could be generated for the picture. This might yield the sentence "A pretty woman in blue hits a man." Bottom: A sentence "A pretty woman in blue hits a man" and the corresponding hierarchical construction structure that TCG can associate with the SemRep (see Section 4 and Fig. 6 for details of the production process).

Consider the specific scene and SemRep shown in Fig. 3. Here, agents and objects are represented as nodes, but we also use nodes to represent attributes. Both nodes and relations may be labeled with "conceptual structures". The properties of a recognized object are attached to the node for that instance of the object, and the semantics of an action are attached to an action relation. Some attached concepts will later be translated into words by the language system (Section 3). 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. However, some of these processes may be directly perceptual (i.e., generated directly by the visual system) while others may be more inferential. It has been claimed (Gallese and Goldman, 1998) 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.

With these analyses, we can provide *a preliminary* definition of SemRep — and we stress again that the same scene can have many SemReps:

2.5. *Definition*

A *SemRep for a scene* consists of a graph – i.e., a set N of nodes, and a subset E of N × N of edges – where the set N of nodes is partitioned into a set O of object nodes and a set A of attribute nodes:

i) Each object node n is linked to a spatial subregion r(n) of the represented scene and is associated with a schema which interprets that region.
ii) Each attribute node is linked to a single object node, and is labeled by a concept which describe a "parameter range" for the schema associated with the object node (e.g., attributes associated with a node representing a person named John might be "smiling" or "tall").
iii) Each edge between object nodes is labeled by a "relation" (which could be a spatial relation, an action, a dominance (is-part-of) relation, or some other relation) which applies to the nodes as currently labeled.
iv) Nodes in SemRep may also be given a *significance* value which expresses the importance of a particular aspect of the scene.

Thus we view SemRep as providing a graphical structure which encompasses one analysis which captures a subset of

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

the agents, objects, actions and relationships that may be present in a given (temporally extended) visual scene.

Thus a spatial relation between nodes m and n actually refers to the regions $r(m)$ and $r(n)$ that they occupy. The subtlety is that we may use the same word to indicate spatial relations within the 2-D plane of the image, or the 3-D world that the image represents.

The *is-part-of* relation is also a spatial relation, but is distinct from those "normal" spatial relations between disjoint objects or regions in the scene. In this case we have $r(m) \subset r(n)$, but possibly with some further semantic description. 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. This fits in with the observation in schema theory that a perceptual schema may return not only the category of the object but also parameters relevant for interaction with the object — and we might now add, post mirror neurons, relevant for recognizing actions in which the object is engaged.

3. Template construction grammar (TCG)

How should linguistics treat idiomatic expressions like *kick the bucket*, *shoot the breeze*, *take the bull by the horns* or *climb the wall*? Rather than taking their meanings as a supplement to general rules of the grammar, Fillmore et al. (1988) 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 (Croft and Cruse, 2005). Here, *constructions* are form-meaning pairings which serve as basic building blocks for grammatical structure — each providing a detailed account of the pairing of a particular syntactic pattern with a particular semantic pattern. Constructions, like items in the lexicon, thus combine syntactic, semantic and even in some cases phonological information.

We are currently implementing a parsing for our own version of construction grammar, Template Construction Grammar (TCG), and have clear ideas on how to extend the work to sentence comprehension (Section 5.3). In some sense, TCG may be seen as a variant of Fluid Construction Grammar (FCG; De Beule and Steels, 2005) with the distinction that TCG grounds its approach to language by using SemRep as the format for its semantics, whereas FCG adopts predicate structure for representing constructions and their (syntactic and semantic) constraints. Moreover, FCG uses logical deduction as the basis for processes of comprehension and production. Another approach which, like TCG and FCG, seeks to place Construction Grammar in a computational framework related to an agent's interaction with the external world is Embodied Construction Grammar (Bergen and Chang, 2005) but no active development of ECG seems to have occurred since 2003, whereas work on FCG continues at an impressive pace. Section 5.1 offers further comparisons of TCG with ECG and FCG.

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.

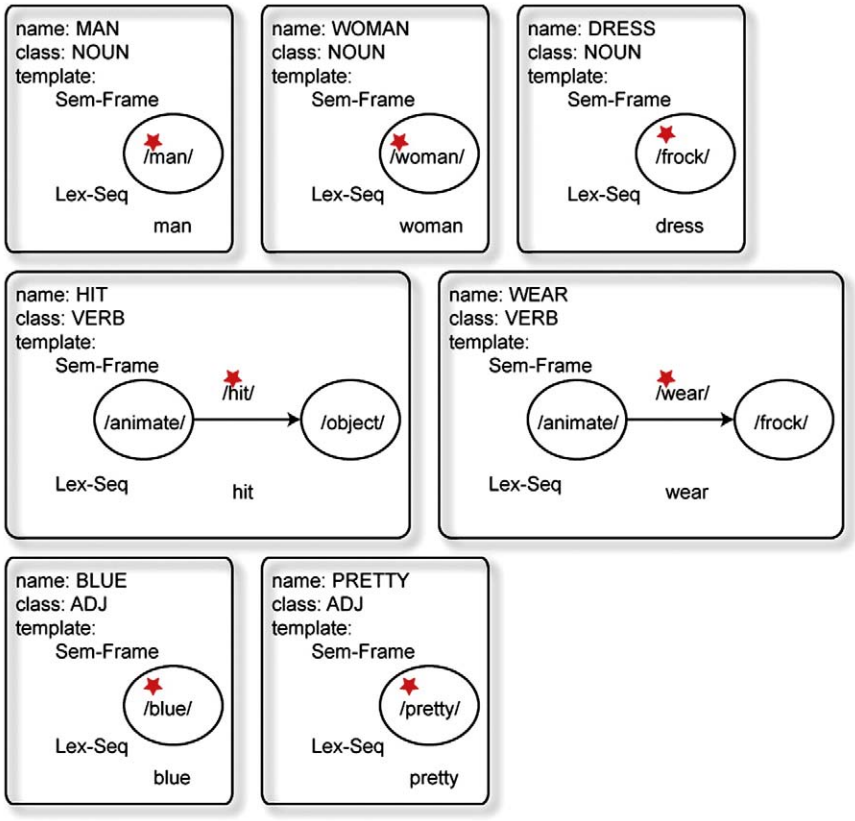

Fig. 4 – Examples of constructions that correspond to elements in the lexicon.

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

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 yield one or more sentences as TCG finds ways to "cover" 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. In production mode, the template acts to match constraints for selecting proper constructions by being superimposed on the SemRep graph. The semantic constraint of each construction is considered to be encoded in the template since the template specifies 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. In the present version of SemRep, we ignore the specification of phonological structure, and ground processing in the deployment of constructions which specify lexical entries of the kind shown in Fig. 4. These constructions show how to associate a word with a particular node. The word shown below the node in each construction in the Lex-Seq section is an actual word of the language (in this case, English). However, what looks like a word inside the node in each construction is actually a "concept-label" which indicates a concept for which the associated word may, but need not be, associated. For example, the concept /frock/ could be associated with the word DRESS, and could also be associated with more general words like CLOTHING. The best way to design constructions to handle such many-to-many relations between word and concept is still under investigation.

We should also note that the types here are standard syntactic categories like NOUN, ADJ (adjective) and VERB. More generally, though, the types may be semantic categories, and these may be either rather general, or may be specific to just a few constructions. Indeed, although we do not use this option in this paper, the type value may actually be a set of actual types, so that the result of applying the construction may fill any slot in any construction for which one of those types is specified. Fig. 5 provides examples of higher-level constructions in which there are slots whose types must be specified. With these examples before us, we can give the following general definition:

3.1. *Definition*

A *construction* is defined by a triple (name, class, template) where:

Name is the name of the construction. It is not involved in the language process – it is only for a reference purpose.

Class specifies the type (e.g., syntactic category) of the result of applying the construction. It determines for which

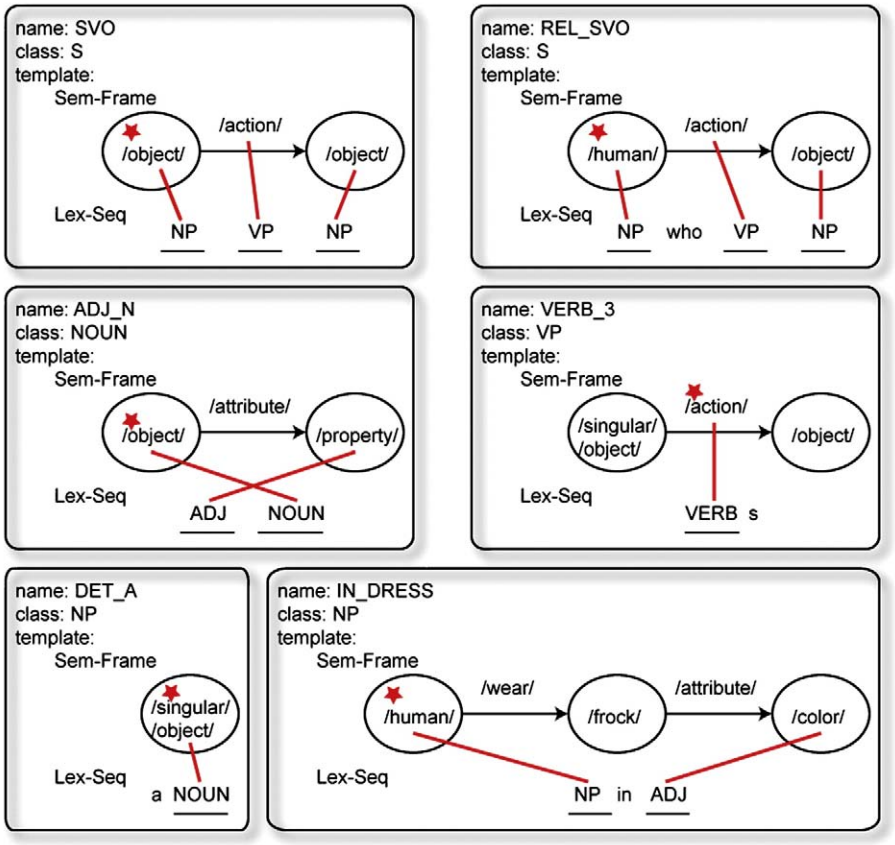

Fig. 5 – Higher-level constructions used to encode grammatical information. Each construction is a SemRep-like graph with either generic or specific labels on the edges and nodes, with each linked to a text or an empty slot. For each slot there may be restrictions as to what can serve as slot fillers.

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

other constructions the result of applying this construction could serve as an input.

The *template* defines the form-meaning pair of a construction and has two components:

- *Sem-Frame* (SemRep frame) defines the meaning part of the construction. It is a part of a SemRep graph that the construction will 'cover' as for its meaning. Each element of this graph is attached with a concept and an activation value as is a typical SemRep graph element. Added to that, Sem-Frame also specifies the 'head' element which acts as a representative element of the whole construction when forming hierarchy with other constructions.
- *Lex-Seq* (lexical sequence) defines the form part of the construction. It is a string of words, morphemes or empty slots. Each slot can be filled with the output of other constructions. Each empty slot specifies the class of a construction that will fill it and the link to the element of Sem-Frame connected to the slot. Only constructions of the same class whose head element is the same as the linked one can be filled in.

Although activation value is not considered here, it can be important in determining the sentence structure, e.g., whether an active or passive sentence is used to describe a scene. 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

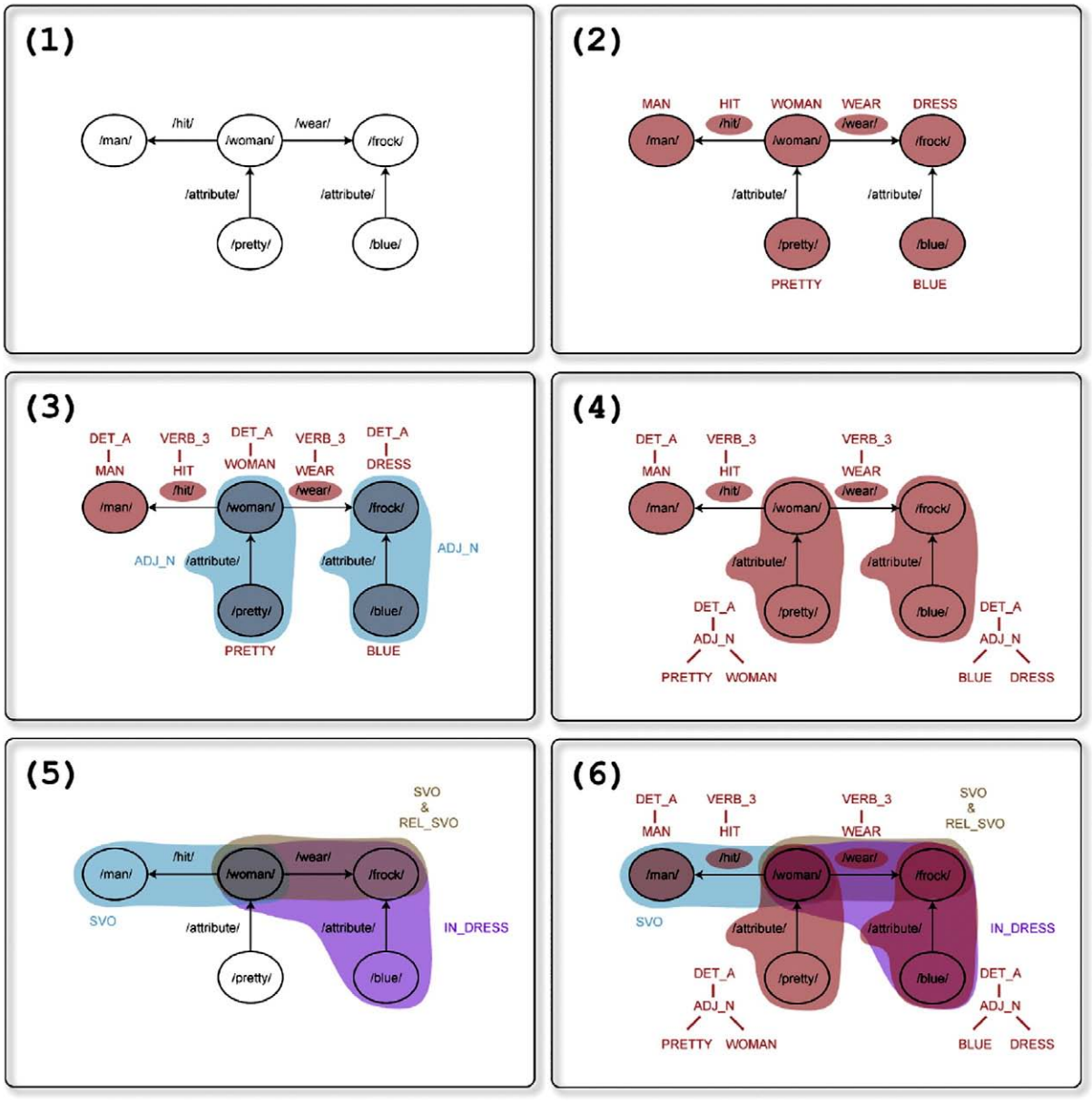

Fig. 6 – Stages in generating a sentence by finding constructions to hierarchically cover a SemRep. Stages (1) through (6). Stages in generating a sentence by finding constructions to hierarchically cover a SemRep. Stages (7) through (12).

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

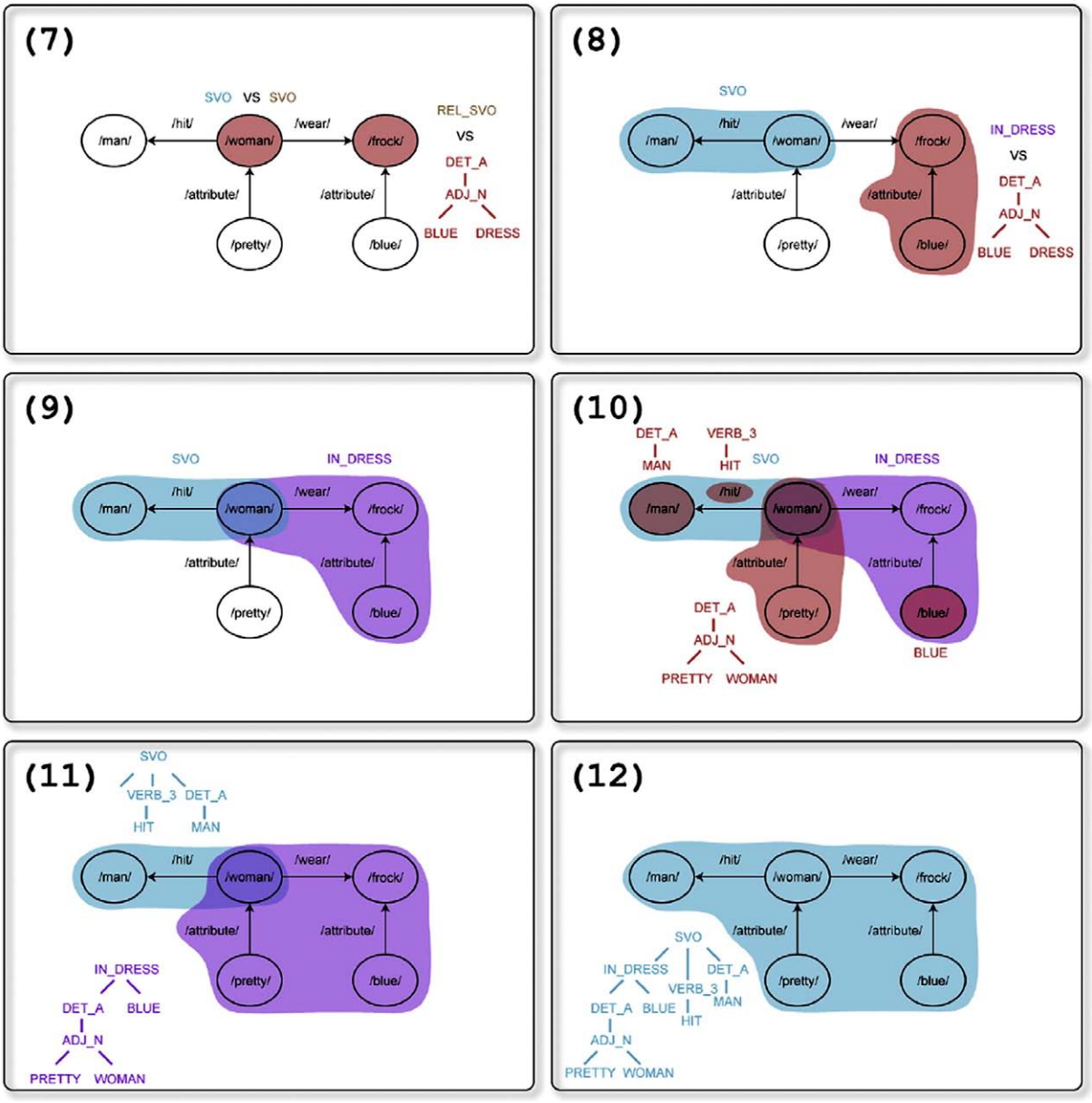

Fig. 6 (*continued*).

is eligible for the conjugation specified in the construction and this grammatical constraint is set by adding negation attributes.

As mentioned above, the template is an (abstract) fragment of a 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 are 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 relate 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. For example, the IN_DRESS construction (Fig. 5) can cover /woman/-/dress/-/blue/ in Fig. 3 even though IN_DRESS does not specifically involve /woman/, /dress/ or /blue/ inside its definition, but only general components such as /human/ or /color/. Note that a concept /human/ is used as a categorical constraint in this case. This allows the system to have great flexibility in producing sentences. A number of constructions with any appropriate concepts available at that moment (e.g. for /human/ concept, /animal/, /mammal/ and /animate object/ are all appropriate) can be selected as candidates.

4. A detailed example: From SemRep to sentence

We now return to the scene and SemRep of Fig. 3, and show (in Fig. 6) the stages of processing the SemRep to yield a description of the scene. In what follows, we assume that

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

the vision system has assigned greatest interest to /woman/ with /man/ next and /blue/ third. /pretty/ and /frock/ would be of about the same importance. Such "top-down" weighting of nodes corresponds to the notion of "focus". Thus if we were looking only at the SemRep for /man/-/hit/-/woman/, focusing on (giving greater weight to) /woman/ would invoke the active construction to yield "A woman hit a man", whereas a focus on man would invoke the passive construction to yield "A man was hit by a woman."

The SemRep is reproduced in Fig. 6(1). In Fig. 6(2), we see the result of applying lexical constructions (recall Fig. 4) to each node and edge as the first step. These not only label words and edges with words, but also assign a category along with the word. The basic class (ADJ, NOUN, VERB) constructions (e.g. for the words DRESS, WEAR, PRETTY, etc.) are evoked first. More complex ones, such as SVO, cannot be evoked at this stage because they have slots to be filled and the components on the SemRep graph have not yet been specified in a way that lets them meet the criteria for filling these slots. At each stage, evoked constructions will be tested against each other with the survivors "attached" to the SemRep graph. In Fig. 6, most panels show "winners" at each round , but Figs. 6 (5), (7), (8), and (9) are in fact intermediate steps to clarify the internal mechanism. In general the SemRep will accumulate alternative constructions with varying "confidence levels" which may rise and fall through successive rounds of competition and cooperation until the final linguistic form is settled upon. Recall that labels like /frock/ on a node and /wear/ on an edge of the graph are *not* words. Instead they denote concepts which, in the final sentence, may or may not be indicated with the word X used in the concept-label /x/. Thus, constructions for the words DRESS, CLOTHING, OUTFIT and FROCK may compete for the /frock/ node and Fig. 6(2) shows the /frock/ node decorated with the word DRESS.

Once all the nodes and relations are initially "covered", ADJ_N constructions are evoked to yield Fig. 6(3), embracing the previously evoked (and attached) NOUN and ADJ class constructions. (Of course, there is no need in general for all these constructions to be invoked at the same stage.) Note that the previously evoked NOUN and ADJ class constructions (e.g. WOMAN, DRESS or BLUE, etc.) are not competing with ADJ_N constructions even though their covering regions are overlapping. This is because ADJ_N class constructions have empty slots for ADJ and NOUN class constructions and they can "fill into" the slots (cooperation). DET_A constructions are evoked as well. In this case, again, they don't compete with (the combined) ADJ_N constructions since DET_A constructions have slots for NOUN class constructions and ADJ_N constructions are of NOUN class, and thus can fill slots in DET_A constructions. We thus see DET_A alone applied to /man/ since it has no linked attributes, whereas we have the hierarchical application (shown in blue in Fig. 6(3)) which yields local parse trees in Fig. 6(4)) applying the ADJ_N and DET-A constructions to both /woman/ and /frock/.

Fig. 6(5) shows what happens as the TCG production process begins to evoke more complex constructions with more slots (note that other constructions are hidden for clarity here – Fig. 6(6) shows the full set). /man/-/hit/-/woman/ will evoke the SVO construction (blue) and /woman/-/frock/-/blue/ will evoke the IN_DRESS (purple) construction. Interestingly, however, /woman/-/wear/-/frock/ will evoke SVO and REL_SVO (brown) at the same time. According to the construction set, /man/-/hit/-/woman/ could also evoke REL_SVO, but we do not show it here since the REL_SVO construction is not evoked with a strong enough confidence level to remain above threshold in the competition with the SVO construction to cover /man/-/hit/-/woman/. Note that the processes demonstrated in Fig. 6(5) through Fig. 6(10) are shown as serial in this exposition, but would actually run in parallel.

Since the SVO for /man/-/hit/-/woman/ (blue) and the SVO for /woman/-/wear/-/frock/ (brown) overlap on the /woman/ node, they will compete (Fig. 6(7) — all other constructions are hidden for clarity). (An alternative is that these two constructions will survive in generating a scene description with at least 2 sentences, but we will not pursue this alternative in the present example.) Meanwhile, the REL_SVO (brown) and DET_A (red) constructions will compete since they are all NP class constructions covering the same node, /frock/.

Fig. 6(8) shows the situation after SVO for /man/-/hit/-/woman/ (blue) wins over SVO for /woman/-/wear/-/dress/ (green) because /man/-/hit/-/woman/ is more active than /woman/-/wear/-/dress/ in the SemRep graph, resulting in the higher similarity value for the former SVO construction. DET_A wins over REL_SVO for the same reason, higher similarity. But note that in this case, cooperation plays a role – the combined constructions (e.g. ADJ_N, BLUE, DRESS) have lent a hand to DET_A.

REL_SVO's similarity
= similarity values of WOMAN construction + WEAR + DRESS.

DET_A's similarity = similarity values of DRESS + BLUE + ADJ_N

In each case, the higher-level construction's similarity is the sum of that for all of its lower level constructions. It applies not only to the attached constructions but also for the newly evoked ones: When entering into competition, similarity scores for the evoked and previously attached constructions are calculated based on the assumption that the slots are all filled in. So REL_SVO is not attached here yet and only DET_A is attached. But since REL_SVO is evoked, there must be appropriate constructions that could fill into REL_SVO at that moment – the similarity has to be calculated as the sum of all component constructions in the hierarchy structure. In fact, if the node /frock/ or /wear/ had been sufficiently stronger relative to /blue/, then REL_SVO would have been the winner, yielding a sentence structure something like "A woman who wears a blue dress…", but here we assume that /blue/ node is more focused on (more 'salient' in the vision term).

To understand what happens next, we need to look at the IN_DRESS construction in Fig. 5 which is a particular application of a more general IN_CLOTHING construction which represents the fact that in English, we may replace "the person who is wearing a blue item of clothing" by "the person in blue" (and similarly for any other color). Fig. 6(9) shows the case in which DET_A loses out in competition with the IN_DRESS construction since they overlap at /frock/-/attribute/-/blue/, their classes are all NP, and the color is more relevant than the nature of the woman's clothing. (All other constructions are hidden for clarity.) Other constructions previously attached in the region,

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

such as DRESS (previously connected to DET_A) and WEAR would compete with IN_DRESS (note that IN_DRESS has no slot for NOUN or VERB types) will die out as well. Only BLUE survives to fill the COLOR slot in IN_DRESS. Fig. 6(10) shows all constructions attached to the SemRep so far. The winner SVO construction now embraces the attached constructions in its region (shaded in blue in Fig. 6(11)), building a hierarchical structure (but note that one slot is still left to be filled). IN_DRESS also embraces the attached constructions and builds a hierarchical structure. Fig. 6(12) shows the final construction hierarchy resulting from the production process. IN_DRESS is used as a filler in SVO, forming a unified structure. Note that even though, Fig. 6(11) and Fig. 6(12) are shown as separate steps, their processing is done simultaneously in the parallel implementation, so the unfinished SVO structure in Fig. 6(11) is not meaningful. Finally, a simple tree traversal on the tree of Fig. 6(12) yields the sentence "A pretty woman in blue hits a man."

As we see, the whole output sentence structure can be very sensitive to a subtle change in activation values of edges or nodes of the SemRep graph. As a result, sentence structure is decided in a flexible manner. The hierarchical structure underlying the sentence is, thus, mainly decided by the production process and does not need to be explicitly encoded in SemRep.

5. Discussion

This concluding Discussion has 3 parts: First, we offer new insight into Template Construction Grammar (TCG) by comparing it with other models of CG, specifically Embodied Construction Grammar (ECG) and Fluid Construction Grammar (FCG), with particular emphasis on the importance of using SemRep for semantic representations in TCG. Second, we offer some pointers to data on neural correlates of vision and language processing relevant to future work in which we will develop TCG within the context of neurolinguistics. Finally, we show how the relation between SemRep and TCG, which in the present paper has been set forth in the context of language production, should serve well as the basis for future work in language comprehension.

5.1. *Comparison with other models of 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 of SemRep involves a sufficiently general graphical structure that we are confident of its extensibility to other meanings. In SemRep, the semantics of an entity is reduced to a node or edge to which is attached a *concept* while the semantics of a scene is represented by the interactive connectivity between the components of a SemRep graph. Each concept is associated with a perceptual schema whose processing is claimed to be instantiated in neural activities, even though the current work implements cooperative computation through direct simulation of schemas rather than through simulation of the brain's neural networks. The perceptual symbols approach (Barsalou, 1999; Barsalou et al., 2003) has some similarities in that it emphasizes the use of sensory-motor representations by the human cognitive system to ground perceptual symbols. However, where Barsalou et al. set up a dichotomy between states in modality-specific systems and redescriptions of these states in amodal representational languages to represent knowledge, our schema theory places more emphasis on multi-modal integration across sensory and motor systems.

Embodied Construction Grammar (ECG) (Bergen and Chang, 2005) and Fluid Construction Grammar (FCG) (De Beule and Steels, 2005) adopt a symbolic strategy for representing semantics. FCG works on logical predicate structures that define semantic/thematic meanings as well as constructing rules of constructions. Due to the nature of its approach, FCG exhibits relatively complex and unintuitive representations with multiple structure types. Although ECG tries for an embodied approach in language understanding by employing X-schemas (Narayanan, 1997) – simulation processes similar in what each represents to the motor schemas of our schema theory – it is fundamentally symbolic. The semantic meanings of constructions are defined by symbolic schemas with variable pre-defined parameters that can be inherited from and assigned to other schemas and constructions. Although these parameters later act as inputs to the simulation by X-schemas, the analytic process for construction manipulation is done on the level of symbolic schemas, not X-schemas, leaving the model symbolic.

Except for the simplicity of representation, another advantage of the TCG approach is that the category of a concept can be driven simply by assessing shared features among the category's members. For example, /grandmother/ and /Peter Pan/ are all /human/ since they both share the common features (or characteristics) of human beings. But only /Peter Pan/ can be categorized as /male/ or /boy/. This makes the semantic matching process in TCG more flexible. Since TCG works on SemRep, the meaning part of the form-meaning pairs of constructions in TCG is represented as a part of a SemRep graph, and the semantic/thematic constraints of argument structure constructions are represented as concepts attached to nodes and relations. For a particular SemRep part to be translated, a number of constructions of different abstraction levels can possibly cover the part without requiring categorical conversion of the concepts. The selected constructions compete and would be selected according to various constraints set at that moment, thus allowing multiple sentences to be produced for a single SemRep. However, as noted by Barsalou et al. (2003), this type of concept system needs to have a strong "content-addressable" memory mechanism that enables easy comparison between similar components. ECG and FCG both show this flexibility to some extent.

Moreover, only a single type of construction is defined in TCG regardless of (syntactic or semantic) level since the categorical information need not be explicitly defined inside constructions and a simple matching process is employed. On the other hand, FCG employs various types of constructions that are basically defined as various types of rules. Each defines an exact transformation process between verbal expression and meaning, in order to capture the categorical divergence in the semantic and syntactic hierarchy. ECG is simpler than FCG in its format, but it also draws on different

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

constructional types that are represented as inheritance among schemas and constructions. As do the rules of FCG, the inheritance strategy in ECG is proposed to define a categorical hierarchy in the semantics and syntax of language.

By including empty slots in constructions and allowing overlap among suitable constructions in the graph-covering process, TCG is capable of handling relative clauses while FCG employs its J-operator to handle hierarchy, whereas ECG does not address this issue explicitly.

5.2. *An assessment of neural correlates*

Although the work on the SemRep/TCG framework reported here has addressed the representation of cooperative computation of schemas without assessing their possible neural correlates, a rapprochement with neurolinguistics provides our long-term motivation. We thus devote this section to a review of relevant neural data and our current thoughts on how to make use of it.

The competition/cooperation between constructions to generate a hierarchical sentence structure by slot filling may be done in the prefrontal region including Broca's area and Brodmann's areas 46/9 since the processing requires a fairly large working memory structure (keeping track of the results of applying multiple constructions) and abstract sequence manipulation ability – reminiscent of Ullman's (2004) characterization of the role of procedural and declarative memory in language processing. The modeling by Dominey and Hoen (2006) which provides a neural model of the basal ganglia and its environs which relates a stripped-down version of Construction Grammar to sequence processing. Kemmerer (2006) uses the framework of Construction Grammar to present the major semantic properties of action verbs and argument structure constructions. He further analyzes neuroanatomical substrates of action verbs and argument structure constructions to argue that the linguistic representation of action is grounded in the mirror neuron system.

Given that a construction defined in TCG is basically a form-meaning pair, it is possible that constructions are distributed throughout the brain, especially centered around the perisylvian regions. These regions, which include classical Broca's (BA 45/44) and Wernicke's (BA 22) areas and the left superior areas of temporal lobe, have long been associated with language processing (Arbib and Bota, 2003; Broca, 1861; Kaan and Swaab, 2002; Wernicke, 1874). Pulvermüller (2001) asserts that a "functional web" linking phonological information related to the articulatory and acoustic pattern of a word form is developed around the perisylvian area since the cortical areas controlling face and articulator movements (the inferior motor cortex and adjacent inferior prefrontal areas) and the auditory system (areas around the superior temporal lobe) might develop strong correlation through direct projection pathways between these two areas (e.g. arcuate fasciculus) from the early babbling phase with stimulation by the self-produced language sounds. Pulvermüller also suggests that "word webs" represent words and aspects of their meaning and include neural circuits in the perisylvian areas storing "word form information" as well as circuits in more wide spread cortical areas for processing related perception and action information, "word meaning information". We may see here another perspective on the relations between "the mirror system for words-as-phonological-objects", "the mirror system for actions" and the network of perceptual and motor schemas shown in Fig. 1.

It has been argued that such semantic representations are topographically distributed across brain areas associated with the process of corresponding categorical properties — concepts of animals are mostly associated with the temporal areas where visual properties are stored whereas concepts of tools or actions are correlated to the motor and parietal areas, generally involved in action and tool use (Chao and Martin, 2000; Martin et al., 1996). The neural association of category-specific concept knowledge, especially in the left hemisphere around the perisylvian areas, is addressed by other studies (Damasio et al., 1996; Martin et al., 1996; Tranel et al., 2003) associated mostly with concrete words, as in the lexicon constructions in TCG (Fig. 4). Pulvermüller et al. (2005) claim that left cortical areas for language and action are linked to each other in a category-specific manner, but one must be careful to distinguish semantic representations from phonological representations and to note that constructions like those of Fig. 5 must assemble words of diverse categories whatever the relative localization of their semantics. Kuperberg et al. (2000) reported that the left-inferior-temporal and fusiform gyri are activated during processing of pragmatic, semantic and syntactic linguistic information. They suggest that this region is responsible for constructing higher representation of sentence meaning. Moreover, the involvement of the perirhinal cortex in object identification and its representation formation integrates not only visual but also multimodal attributes (e.g. smell or texture) (Murray and Richmond, 2001). Also, the perirhinal cortex is suggested to be responsible for the relatively long-term memorization of such representations (Buffalo et al., 1998).

The main operational theme of TCG is the selection of constructions by competition and cooperation and building the hierarchical sentence structure by slot filling among those selected constructions. It is hypothesized that Broca's area is mainly involved in this operation in that it is in most part the syntactic manipulation of constructions which is supported by the matching process of construction semantics – i.e. matching semantic meanings/categorical information of construction for deciding winners and arranging them in a certain order. In fact, Broca's area is activated more when handling sentences of complex syntactic structure than of simple structure (Stromswold et al., 1996). It has been further hypothesized that BA 44 of Broca's area is for handling the arrangement of lexical items whereas BA 45 is for retrieving semantic or linguistic components during the matching and ordering process. Note that since the manipulation of the orderly arrangement of phonetic components (both in verbal and sign language) is fundamentally the same as the manipulation of lexical items, BA 44 is assigned the role of processing the lexical sequence (Lex-Seq) of constructions in TCG, which is the unified representation of slots and phonemes together.

Horwitz et al. (2003) conducted a PET study which showed that BA 45, not BA 44, is activated by both speech and signing during the production of language narratives done by bilingual subjects, whereas BA 44, but not BA 45, is activated by the

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

generation of complex articulatory movements of oral/laryngeal or limb musculature. Indeed, it should not be a coincidence that the left-inferior prefrontal regions are a crucial component of the proposed phonological rehearsal circuit (Aboitiz and Garcia, 1997; Smith et al., 1998). Moreover, Corina et al. (1999) reported the selective participation of BA 44 in phonetic aspects of linguistic expressions.

5.3. *Working memory*

A large amount of working memory is required to support the operations mentioned above, both working memory for construction of a visual episode and the associated SemRep, and the working memory for the words and constructions involved in the description of the scene. Working memory for the visual scene may be in the right parieto-occipital area, since patients with lesions in this region have (dorsal) simultanagnosia — they can recognize objects, but cannot see more than one object at a time (Coslett and Saffran, 1991; Michel and Henaff, 2004). Smith et al. (1998) proposed that the storage buffer for verbal working memory roughly corresponds to the left posterior parietal cortex (BA 40) and its connection to the left prefrontal cortex while the executive component for processing the contents of working memory lies in the left dorsolateral prefrontal cortex (DLPFC; BA 9/46). Similarly, monkey prefrontal cortex is involved in sustaining memory for object identity and location (Rainer et al., 1998) and the processes for the object spatial location and the object characteristics are segregated in different regions — the posterior parietal cortex connected to the DLPFC and the connections of the inferior temporal lobe and the inferior convexity of the prefrontal cortex, respectively (Wilson et al., 1993). Aboitiz et al. (Aboitiz and Garcia, 1997; Aboitiz et al., 2006a,b) claim that the projection of phonological representations created in Wernicke's area through the inferoparietal areas to Broca's area forms a phonological rehearsal device as well as a working memory circuit for complex syntactic verbal processes. Arbib and Bota (2006) – especially their Fig. 5.7 – compare this approach with the Mirror System Hypothesis.

5.4. *Comprehension*

Although only production is addressed in this paper, a similar mechanism based on the competition and cooperation framework can be used for comprehension in TCG. In this case, the same set of constructions can be used as well, only with the application direction reversed. FCG also uses the same set of rules and constructions in both directions. However this very flexibility in computational terms fails to address a well-known psycholinguistic fact — that we are often capable of understanding sentences even when we have not mastered the constructions needed to generate them. The resolution of this (whose details lie outside the scope of the present paper) is that we engage a cooperative computation paradigm for perception which does not lie wholly within the linguistic domain. The parsing process may deliver fragments of a SemRep which is incompletely integrated, but then the processes which link SemRep to the schema instantiation processes of visual scene perception can repair the defective SemRep to yield a plausible scene representation. In VISIONS, the schema instance level may invoke the intermediate database to in turn invoke further processing to return answers needed to assist the competition and cooperation between schema instances, so we must understand that SemRep is not an isolated graphical representation but is instead linked to the schema instance level though giving only a partial view of it. Each node is linked to a region of the image either via the schema instance for that region (e.g., for an object) or to parameters describing that region (as when an attribute is linked to a node corresponding to that region) or to the linkage between regions related to agents or objects (as in the case of actions or spatial relations) which may encompass a somewhat larger region. In the same fashion, we note that the linkage of a SemRep to a Schema Instance Map allows the SemRep to be expanded or adjusted as the demands of narration require. This is true in the case of finding information which serves to disambiguate a description, as well as in answering a question about a scene.

Of course, this does not guarantee that the SemRep is "correct" for the given sentence. Such an interplay of linguistic and cognitive processing has been invoked in the study of aphasia. Indeed, Piñango (2006) stresses that comprehension can take place despite syntactic impairment, *but only if the sentence's semantic structure is rich enough*. She relates this to the syntax-independent semantic combinatorial mechanisms of Culicover and Jackendoff (2005), but we would suggest that our SemRep/TCG framework, by employing cooperative computation of schemas, has more promise for a rapprochement with neurolinguistics.

The TCG formalism exploits the combination of attributes or properties for the concept attached to a node or edge in a SemRep graph to compare conceptual entities. During production of sentences, a given graph 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 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.

The VISIONS system provided our motivating example of how to build a system in which competition and cooperation between schema instances can generate an interpretation of a static visual scene. The HEARSAY speech understanding system (Erman et al., 1980) provides a cooperative computation view of sentence parsing/interpretation which operates in the time domain, proceeding from the spectrogram for a spoken sentence to a possible syntactic analysis and semantic interpretation of the utterance. 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 a single coherent symbolic representation of the syntax and semantics of the most plausible interpretation of the auditory input. HEARSAY was implemented on a serial computer, and the

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

designers thus had to put much effort into the design of algorithms for the serial scheduling of "knowledge sources", each of which corresponds, in our terminology, to the application of one from a group of related schemas. Arbib and Caplan (1979) discussed how this serial architecture might be converted into a "neuro-HEARSAY" based on the competition and cooperation of schemas in the brain (see also Arbib et al., 1987). This neuro-HEARSAY provided one inspiration for the present work, and future work must move beyond it in developing neural models of the interaction of SemRep and TCG in the process of providing the semantics (SemRep) of a given utterance (sequence of words).

REFERENCES

Aboitiz, F., Garcia, V.R., 1997. The evolutionary origin of the language areas in the human brain. A neuroanatomical perspective. Brain Res. Brain Res. Rev. 25, 381–396.

Aboitiz, F., García, R., Brunetti, E., Bosman, C., 2006a. The origin of Broca's area and its connections from an ancestral working memory network. In: Grodzinsky, Y., Amunts, K. (Eds.), Broca's region. Oxford University Press, Oxford, pp. 3–16. Vol.

Aboitiz, F., Garcia, R.R., Bosman, C., Brunetti, E., 2006b. Cortical memory mechanisms and language origins. Brain Lang. 98, 40–56.

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

Arbib, M.A., 2003. Rana computatrix to human language: towards a computational neuroethology of language evolution. Philos. Transact. A Math. Phys. Eng. Sci. 361, 2345–2379.

Arbib, M.A., 2005. From monkey-like action recognition to human language: an evolutionary framework for neurolinguistics (with commentaries and author's response). Behav. brain sci. 28, 105–167.

Arbib, M.A., 2006. Aphasia, apraxia and the evolution of the language-ready brain. Aphasiology 20, 1–30.

Arbib, M.A., Bota, M., 2003. Language evolution: neural homologies and neuroinformatics. Neural Netw 16, 1237–1260.

Arbib, M.A., Bota, M., 2006. Neural homologies and the grounding of neurolinguistics. In: Arbib, M.A. (Ed.), Action to language via the mirror neuron system. Cambridge University Press, Cambridge, pp. 136–173. Vol.

Arbib, M.A., Caplan, D., 1979. Neurolinguistics must be computational. Behav. brain sci. 2, 449–483.

Arbib, M.A., Didday, R.L., 1971. The organization of action-oriented memory for a perceiving system. I. The basic model. J. cybern. 1, 3–18.

Arbib, M.A., Lee, J., 2007. Vision and action in the language-ready brain: from mirror neurons to SemRep. In: Mele, F. (Ed.), BVAI 2007 (Brain Vision & Artificial Intelligence, 2007), LNCS 4729. Springer-Verlag, Berlin, pp. 104–123. Vol.

Arbib, M.A., Liaw, J.-S., 1995. Sensorimotor transformations in the worlds of frogs and robots. Artif. Intell. 72, 53–79.

Arbib, M.A., Conklin, E.J., Hill, J.C., 1987. From Schema Theory to Language. Oxford University Press, New York. Vol.

Arbib, M.A., Érdi, P., Szentágothai, J., 1998. Neural Organization: Structure, Function, and Dynamics. The MIT Press, Cambridge, MA. Vol.

Barsalou, L.W., 1999. Perceptual symbol systems. Behav. Brain Sci. 22, 577–609 discussion 610–60.

Barsalou, L.W., Kyle Simmons, W., Barbey, A.K., Wilson, C.D., 2003. Grounding conceptual knowledge in modality-specific systems. Trends Cogn. Sci. 7, 84–91.

Bartlett, F.C., 1932. Remembering. Cambridge University Press, Cambridge. Vol.

Bergen, B., Chang, N., 2005. Embodied construction grammar in simulation-based language understanding. In: Östman, J.-O., Fried, M. (Eds.), Construction grammar(s): Cognitive and cross-language dimensions. John Benjamins, Amsterdam, pp. 147–190. Vol.

Bonaiuto, J., Rosta, E., Arbib, M., 2007. Extending the mirror neuron system model, I : audible actions and invisible grasps. Biol. Cybern. 96, 9–38.

Broca, P.P., 1861. Perte de la parole. Bull. Soc. anthropol. Paris 2, 235–238.

Buffalo, E.A., Reber, P.J., Squire, L.R., 1998. The human perirhinal cortex and recognition memory. Hippocampus 8, 330–339.

Chao, L.L., Martin, A., 2000. Representation of manipulable man-made objects in the dorsal stream. Neuroimage 12, 478–484.

Corina, D.P., McBurney, S.L., Dodrill, C., Hinshaw, K., Brinkley, J., Ojemann, G., 1999. Functional roles of Broca's area and SMG: evidence from cortical stimulation mapping in a deaf signer. Neuroimage 10, 570–581.

Coslett, H.B., Saffran, E., 1991. Simultanagnosia. To see but not two see. Brain 114 (Pt 4), 1523–1545.

Croft, W., Cruse, D.A., 2005. Cognitive Linguistics. Cambridge University Press, Cambridge. Vol.

Culicover, P., Jackendoff, R., 2005. Simpler Syntax. Oxford University Press, Oxford. Vol.

Damasio, H., Grabowski, T.J., Tranel, D., Hichwa, R.D., Damasio, A.R., 1996. A neural basis for lexical retrieval. Nature 380, 499–505.

De Beule, J., Steels, L., 2005. Hierarchy in fluid construction grammar. In: Furbach, U. (Ed.), Proceedings of the 28th Annual German Conference on AI, KI 2005, Lecture Notes in Artificial Intelligence, vol. 3698. Springer-Verlag, Berlin, pp. 1–15. Vol.

Dominey, P.F., Hoen, M., 2006. Structure mapping and semantic integration in a construction-based neurolinguistic model of sentence processing. Cortex 42, 476–479.

Draper, B.A., Collins, R.T., Brolio, J., Hanson, A.R., Riseman, E.M., 1989. The schema system. Int. j. comput. vis. 2, 209–250.

Erman, L.D., Hayes-Roth, F., Lesser, V.R., Reddy, D.R., 1980. The HEARSAY-II speech understanding system: integrating knowledge to resolve uncertainty. Comput. Surv. 12, 213–253.

Fagg, A.H., Arbib, M.A., 1998. Modeling parietal-premotor interactions in primate control of grasping. Neural Netw. 11, 1277–1303.

Fillmore, C.J., Kay, P., O'Connor, M.K., 1988. Regularity and idiomaticity in grammatical constructions: the case of let alone. Lang. Cogn. Processes 64, 501–538.

Gallese, V., Goldman, A., 1998. Mirror neurons and the simulation theory of mind-reading. Trends Cogn. Sci. 2, 493–501.

Goodale, M.A., Milner, A.D., 1992. Separate visual pathways for perception and action. Trends Neurosci. 15, 20–25.

Grice, H.P., 1969. Utterer's meaning and intention. Philos. Rev 78, 147–177.

Head, H., Holmes, G., 1911. Sensory disturbances from cerebral lesions. Brain 34, 102–254.

Hickok, G., Poeppel, D., 2004. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92, 67–99.

Hickok, G., Poeppel, D., 2007. The cortical organization of speech processing. Nat. Rev. Neurosci. 8, 393–402.

Horwitz, B., Amunts, K., Bhattacharyya, R., Patkin, D., Jeffries, K., Zilles, K., Braun, A.R., 2003. Activation of Broca's area during the production of spoken and signed language: a combined cytoarchitectonic mapping and PET analysis. Neuropsychologia 41, 1868–1876.

Itti, L., Arbib, M.A., 2006. Attention and the minimal subscene. In: Arbib, M.A. (Ed.), Action to language via the mirror neuron system. Cambridge University Press, Cambridge, pp. 289–346. Vol.

Kaan, E., Swaab, T.Y., 2002. The brain circuitry of syntactic comprehension. Trends Cogn. Sci. 6, 350–356.

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075

Kemmerer, D., 2006. Action verbs, argument structure constructions, and the mirror neuron system. In: Arbib, M.A. (Ed.), Action to language via the mirror neuron system. Cambridge University Press, Cambridge, pp. 347–373. Vol.

Kuperberg, G.R., McGuire, P.K., Bullmore, E.T., Brammer, M.J., Rabe-Hesketh, S., Wright, I.C., Lythgoe, D.J., Williams, S.C., David, A.S., 2000. Common and distinct neural substrates for pragmatic, semantic, and syntactic processing of spoken sentences: an fMRI study. J. Cogn. Neurosci. 12, 321–341.

Lyons, D.M., Arbib, M.A., 1989. A formal model of computation for sensory-based robotics. IEEE trans. robot. autom. 5, 280–293.

Marr, D., 1982. In: Freeman, W.H. (Ed.), Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Vol., New York.

Martin, A., Wiggs, C.L., Ungerleider, L.G., Haxby, J.V., 1996. Neural correlates of category-specific knowledge. Nature 379, 649–652.

Michel, F., Henaff, M.A., 2004. Seeing without the occipito-parietal cortex: simultagnosia as a shrinkage of the attentional visual field. Behav. Neurol. 15, 3–13.

Murray, E.A., Richmond, B.J., 2001. Role of perirhinal cortex in object perception, memory, and associations. Curr. Opin. Neurobiol. 11, 188–193.

Narayanan, S.S., 1997. Knowledge-based Action Representations for Metaphor and Aspect (KARMA). Engineering: Computer Science. University of California, Berkeley. Vol.

Oztop, E., Arbib, M.A., 2002. Schema design and implementation of the grasp-related mirror neuron system. Biol. Cybern. 87, 116–140.

Piaget, J., 1971. Biology and knowledge: An essay on the relations between organic regulations and cognitive processes [Translation of (1967) Biologie et connaissance: Essai sur les relations entre les régulations organiques et les processus cognitifs. Paris: Gallimard.], Vol.. Edinburgh University Press, Edinburgh.

Piñango, M.M., 2006. Understanding the architecture of language: the possible role of neurology. Trends Cogn. Sci. 10, 49–51.

Pulvermüller, F., 2001. Brain reflections of words and their meaning. Trends Cogn. Sci. 5, 517–524.

Pulvermüller, F., Hauk, O., Nikulin, V.V., Ilmoniemi, R.J., 2005. Functional links between motor and language systems. Eur. J. Neurosci. 21, 793–797.

Rainer, G., Asaad, W.F., Miller, E.K., 1998. Memory fields of neurons in the primate prefrontal cortex. Proc. Natl. Acad. Sci. U. S. A. 95, 15008–15013.

Rizzolatti, G., Arbib, M.A., 1998. Language within our grasp. Trends Neurosci. 21, 188–194.

Schmidt, R.A., 1975. A schema theory of discrete motor skill learning. Psychol. Rev. 82, 225–260.

Smith, E.E., Jonides, J., Marshuetz, C., Koeppe, R.A., 1998. Components of verbal working memory: evidence from neuroimaging. Proc. Natl. Acad. Sci. U. S. A. 95, 876–882.

Stromswold, K., Caplan, D., Alpert, N., Rauch, S., 1996. Localization of syntactic comprehension by positron emission tomography. Brain Lang. 52, 452–473.

Tranel, D., Kemmerer, D., Adolphs, R., Damasio, H., Damasio, A.R., 2003. Neural correlates of conceptual knowledge for actions. Cogn. Neuropsychol. 20, 409–432.

Ullman, M.T., 2004. Contributions of memory circuits to language: the declarative/procedural model. Cognition 92, 231–270.

Wernicke, C., 1874. Der aphasische symptomencomplex. Cohn and Weigert, Breslau. Vol.

Wilson, F.A., Scalaidhe, S.P., Goldman-Rakic, P.S., 1993. Dissociation of object and spatial processing domains in primate prefrontal cortex. Science 260, 1955–1958.

Please cite this article as: Arbib, M.A., Lee, J. Describing visual scenes: Towards a neurolinguistics based on construction grammar. Brain Research (2008), doi:10.1016/j.brainres.2008.04.075