LINKING EYES TO MOUTH:

A SCHEMA-BASED COMPUTATIONAL MODEL FOR DESCRIBING VISUAL SCENES

by
Jinyong Lee

A **Revised Version**[1] of the Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA

DOCTOR OF PHILOSOPHY
In COMPUTER SCIENCE

September 2013

[1] This is a revised version of the original thesis submitted in August 2012. This version of thesis contains one additional chapter (Chapter 3) which contains two new sections (Section 3.3 and Section 3.4). It also contains updated/revised contents in places.

Acknowledgements

I would like to thank Professor Michael A. Arbib for being an excellent advisor, a trusted mentor, and a respectable role model throughout the Ph.D. program in USC. I have been mostly benefited from his insight and knowledge on the nature of human and primate cognitive system.

I would also like to acknowledge the constant support and encouragement from my valuable lab mates, Victor Barrés, Jimmy Bonaiuto, Brad Gasser, Rob Schuler, Arthur Simons, and Nader Noori (in iLab). I am thankful for exciting discussions and enjoyable research atmosphere shared with all of my colleagues.

I also like to show my gratitude to Dr. Laurent Itti for allowing me to use the eye-tracking equipment with technical support and to Dr. Elsi Kaiser for her valuable advice throughout the thesis writing.

This thesis would not have been possible without patient support from my family, my mother Jungsim Boo and my brother Jinwon Lee. Most of all, I would like to give my biggest thanks to my fiancé Eunyoung Song, who offered her constant care, love, and trust.

Table of Contents

Abstract

The present thesis 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. As the first step, we use the task of describing visual scenes to explore the suitability of the currently proposed framework of a schema-based linguistics. We developed a new kind of semantic representation, *SemRep*, which is an abstract form of visual information with an emphasis on the spatial linkage of entities, attributes and actions. SemRep provides a compact graph-like structure with enough formal semantics for verbal description of a scene, reducing the relatively complex task of semantic processing to a graph matching task. The present thesis reports 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. Constructions, represented as schema instances in our approach, compete and cooperate to cover the SemRep to produce a description of the visual scene at hand. In our approach, the vision system interprets a part of the scene under attention by creating or updating the corresponding SemRep while the language system applies constructions on that part of SemRep by the principles of TCG. The current work proposes specific mechanisms on how a representation (i.e. SemRep) is built from the perceived visual scene and what influences the choice of constructions for the produced utterances. More specifically, the complexity of a perceived event and the constraints on available computational resources are hypothesized to be the main driving force of the resultant sentential structure being produced. The former affects the coverage of the perceived *subscene*, which represents a particular view on the scene at a certain moment, and the resultant formulation of SemRep. The latter, which is parameterized as the *threshold of utterance*, limits the amount of time and constructions used for formulating descriptions, resulting in different degrees of "well-formedness" of produced sentences. To test hypotheses, we conducted a series of eye-tracking experiments with experimental settings to induce various levels of event complexity (e.g. showing scenes of different event structures) and threshold (e.g. imposing time pressure). Based on the examination on the time-locked eye movements and recorded speech, the present thesis presents supporting evidence for the proposed mechanisms. We conclude by demonstrating how the combinations of various levels of threshold and event complexity in the framework of SemRep and TCG can address both of the apparently opposing strategies in sentence production: the "structural view" which asserts the preparation of sentential structure and the preparation of each constituent are interleaved, and the "incremental view" that claims that those are separated in an orderly fashion.

Chapter 1. Introduction

The overall aim of this thesis is to provide clues on how linguistic processes relate to mechanisms of visual perception and to propose an integrated framework of those two systems. As an initial effort to do this, the current work describes a computational approach to the perception of visual scenes and the production of the scene description. We propose "SemRep" as a unique form of semantic representation of a visual scene. We also provide the detailed descriptions of a theoretical framework and an implemented model of utterance production, Template Construction Grammar (TCG). We propose SemRep as a type of graph-like representation that bridges between the vision and language system while TCG as a conceptual model of the language (production) system, which runs on input from the vision system, provided as SemRep. TCG is distinguished by its explicit usage of schema theory and its computational paradigm – constructions are regarded as schema instances that compete and cooperate with each other to converge on the solution, which is in our case the description of a perceived scene.

The initial framework of our approach was described by Itti and Arbib (2006). They discussed how perception of a "minimal subscene" associating an agent and an action to one or more objects may underlie processes of scene description and question-answering, linking the schematic structure of visual scenes to language structure – the question that dates back to the "two visual system" model of visual perception (Didday & Arbib, 1975). Knott (Knott, 2003) proposed a model based on a very similar framework – in his model, the sensorimotor sequence of attention to the scene (the scanpath) is translated directly into the operations involved in constructing the syntactic tree for its description. However, his approach differs from our approach of TCG in that it adopts a version of the Minimalist approach (Chomsky, 1995) where the clause syntax is directly mapped onto the sensorimotor model of action perception and execution. Moreover, the scene description is built on the eye movements rather than the state of the symbolic WM since sensorimotor sequences are directly linked to sentences. This is in contrast with TCG, which exhibits more capability in processing complex sentential structures (with recursion) by its explicit usage of SemRep as the symbolic WM.

Furthermore, we designed and conducted a series of eye tracking experiments to explore possible explanations and solutions to linking of visual perception and utterance production. Analysis of the experimental data provides findings and observations to test the hypotheses that we have proposed during the development of SemRep and TCG. More specifically, we discuss experimental evidence for supporting the notion of "subscene" and the mechanisms through which a subscene is perceived and encapsulated into a SemRep. Moreover, we address the supporting evidence for the threshold of utterance, which we propose as a theoretical construct that sets an upper bound on available computational resources during utterance production. Experimental evidence for the principles of utterance production, which are designed to capture the dynamics of eye fixations and the related utterances in relation to the different levels of threshold, are also discussed.

The current thesis consists of four main chapters. Chapter 2 provides the theoretical background and the list of the relevant literature to SemRep. We do not model the specific processes of building a SemRep, but instead focus on its role as an internal representation of a perceived scene on which linguistic processes work to produce scene description. We also provide an account of the neural processes in the vision system whereby a SemRep may be generated and the

neurophysiological and behavioral evidence that supports different aspects of the information encoded in a SemRep, which serves as input to the system of TCG, described in the later chapter.

In Chapter 3, we discuss three types of cortical network: semantics and concepts, visual perception, and linguistic process. All of these networks are described in terms of a working memory (WM) network, each of which takes a distinctive role in the current framework of visual scene description. While thoroughly reviewing neurophysiological and behavioral evidence relevant to each of these networks, we conclude the chapter by proposing an integrative WM framework of these three networks – the Visuo-Linguistic Working Memory (VLWM). The VLWM encompasses cortical structures relevant to performing the task of visual scene description while providing a "shared workspace" for the interplay between visual and linguistic (as well as semantic) processes during the task.

Moreover, Chapter 4 provides a detailed account on TCG while highlighting how the compact, yet formal, structure of SemRep is exploited by a schema-based approach. The theoretical framework of TCG and the implementation details (with simulation results) of TCG are provided in a separate manner in order to emphasize the generality of TCG in its application to the process of scene description production. In this chapter, we relate TCG to the analysis of how descriptions of a scene may vary under various constraints, such as time pressure, by testing the model with different levels of threshold that limits computational resources spent in formulating utterances. We also propose principles of utterance production during the task of scene description, which we claim provides explanations for a number of particular patterns observed from speakers' performance on utterance production during scene perception.

Lastly, Chapter 5 reports two eye-tracking experiments with detailed analysis results. We designed and conducted a series of eye-tracking experiments to test our hypotheses on how semantic representation is built from acquired visual information and how it influences the produced utterances. Especially, the analysis results focus on providing evidence on the validity of the threshold of utterance and its influence on the "well-formedness" of produced utterance. With emphasis on various experimental circumstances that could induce different levels of threshold, we propose that the combination of threshold and the principles of utterance production proposed in TCG may provide a reconciliatory explanation for the two seemingly opposing views in the studies of scene apprehension and linguistic formulation: (1) the incremental view, which claims that perceptual salience affects sentence production, as reflected by the order of mention or structure of the sentence, and (2) the structural view, which argues that perceptual saliency does not play a significant role in utterance order as holistic apprehension of scenes precedes formulation of sentences. We conclude the chapter by suggesting that those two views are not the results of two mutually exclusive mechanisms, but outcomes of two extreme cases generated from a single mechanism (as delineated by SemRep and TCG) with a change of policy.

Chapter 2. Schema-based System of Scene Perception

2.1. Schema Theory

In his influential book “The Organization of Behavior (1949)”, Hebb argued that higher brain processes are realized as functional units above the level of the neuron. He regarded the basis of cognition as the interference patterns of either individual neurons or their mass activity in the entire cortex. In the present work of linking vision and language, our approach is based on a similar, yet more computationally specified, framework – a version of *schema theory* in which *schemas* are a type of distributed program that captures functions of neural networks in the brain (Michael A. Arbib, 1981; Michael A. Arbib, Érdi, & Szentágothai, 1998) or else provides units for distributed versions of technological computations. We must stress that a schema as a functional unit should not be equated with a structural unit (e.g. a segregated neural circuit in the brain). In general, a single schema may be implemented across several structural units while a single structural unit may contribute to several schemas (Michael A. Arbib & Liaw, 1995). Schema theory is designed to provide a symbolic level of computational modeling that aims at facilitating later transfer to neural level implementation. We therefore can rest content with functional models which yield the patterns of behavior of the animal or human as seen “from the outside”. We can also probe further and restructure our models in the light of lesion studies or brain imaging data (or single-cell recording in the case of animal studies) to help us understand how this behavior is mediated by the inner workings of the brain.

In this approach to schema theory, schemas are generally defined in three types: *perceptual schemas*, which are for recognizing objects or states of the world, *motor schemas*, which are for interactions with those objects and states perceived by perceptual schemas, and more abstract *coordinating schemas*, which are for mediating and coordinating schemas and their interactions. Perceptual schemas can be coupled with motor schemas to form (possibly mediated by coordinating schemas) *coordinated control programs* (Michael A. Arbib, 1981), an assemblage of schemas 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 (Michael A. Arbib, 2002).

Moreover, the notion of schema is “recursive”. A schema defined functionally may be analyzed as a coordinated control program of finer schemas, and so on until such time as a secure foundation of neural localization or technological implementation is attained. A schema is also “learnable”. New schemas may be formed as assemblages of old schemas, but once formed, a schema may be tuned by some adaptive mechanism, which allows them to start as composite but emerge as primitive, much as a skill is honed into a unified whole from constituent pieces (Michael A. Arbib, 1995). When we learn how to peel an apple, for example, we may quickly approximate the skill by marshaling a stock of existing schemas, such as rotating an object or skinning with a knife, and then tune the resultant assemblage through experience to emerge with a new schema for skilled performance of the task. These examples illustrate the case of motor schemas, but imply improvement in perceptual schemas that provide data to guide the actions; a similar case can be made for perceptual schemas too. A perceptual schema may consist of a number of subschemas that capture different aspects of the perceived entity – e.g. the perception of an apple may involve a number of perceptual schemas for its shape, texture, size, and smell – and a new perceptual schema might emerge from a set of existing perceptual schemas that are grouped together over experience – e.g.

an experienced farmer may develop a perceptual schema for the specific species of an apple, which generally requires a number of other perceptual schemas for the specific shape, texture, or color of an apple.

Given that one of the main purposes of the present work is to establish the framework for the coordinated mechanisms of perceiving a scene and producing a verbal description thereof, perceptual schemas are of our main interest. Perceptual schemas can serve to pass parameters describing the state of the world to motor schemas which will control the agent's interaction with the world. For instance, for a given motor schema of "peeling" an apple, the perceptual schema for the apple not only specifies the perceptual properties of the apple, such as color, or smell, but also specifies the parameters relevant for the motoric action of peeling, such as the size of the apple for grasping, or the hardness of the peel. Thus, in the schema-theoretic approach, the perception of an apple is not mere categorization of the apple as an "apple", but may provide access to a range of parameters relevant to interaction with the apple at hand (Michael A. Arbib, 2002).

However, one should note that recognizing an apple may be linked to many different courses of action, such as "to place the apple on a table", "to choose the apple in a market", "to peel the apple", or "to eat the apple". Each particular action requires a different set of parameters that should be delivered from perception of the apple – some may only need more generic parameters, such as size or shape, while others may need parameters more specific to the apple, such as ripeness or species – and this may involve a number of perceptual schemas each of which captures a different aspect of the apple – some may capture only the size of any spherical objects, some may capture smell of any fruit, and some may categorize the species of an apple (only for a skilled farmer), etc. Therefore, there is no one "grand apple schema" which links all "apple perception strategies" to "every act that involves an apple", but rather the "perception" of an apple is defined in terms of a set of perceptual schemas that are invoked in necessity to provide the relevant information to the particular motor schemas chosen under the current plan of action.

Thus, schema-based perception is *action-oriented* (see Michael A. Arbib, 1972 for more detailed description on the framework of action-oriented perception). Action-oriented perception asserts that the perception process should not be considered as a passive act, but as an active process through which the perception system continuously interprets the sensory experience in terms of performing the action goals set by the organism. However, in describing a scene, few of the parameters that guide action need enter conscious awareness, and so we will specifically address the case where a parameter is "promoted" to an explicit attribute which can form part of the description.

Schema-based modeling of perception emphasizes the role of a visual working memory updating a *schema assemblage* (Michael A. Arbib, 1989) that combines the *schema instances* encoding relevant aspect of, and plans for interaction with, the current environment. This assemblage is dynamic, as certain schema instances are discarded from memory ("de-instantiated" or "eliminated") while others are added ("instantiated" or "invoked"). Long term memory, which defines the world knowledge of the organism, provides the stock of schemas from which a schema assemblage may be assembled.

Therefore, perception of a scene may be modeled as invoking instances of perceptual schemas for certain aspects of the scene rather than simply tagging labels to the presented elements of the scene. Once the perception of an object has been made, a separate schema instance is created in one's working memory for representing each instance of the object. Each schema instance is tuned with appropriate parameters to represent the particularities of the object it represents (though, as noted above, only a few of these may be "promoted" into the verbal description). Thus, a schema instance acts as an "active

copy" since a schema instance is a "parameterized" version of the base schema in that it represents specific configurations of the object that are represented by more general terms in the base schema which act as the "master copy".

For example, in a scene where there are three chairs, three "chair" schema instances are instantiated from the perceptual schema of a chair, and the particularities of each chair instance (e.g. the orientation of the chair, the size of the seat, or height of the arm rest, etc.) are encoded as the parameter values in each schema instance. These schema instances, created in the working memory, form a schema assemblage as an internal representation of the perceived scene so far. Each schema instance also has a *confidence level* so that during the early stages of scene recognition, alternative schemas can compete to form part of the final interpretation.

Note that in the present thesis, we will use something akin to a tagged spatial structure as the bridge from vision to language. Thus, the key for future research is that the language system to access the dynamics of visual perception via this intermediary.

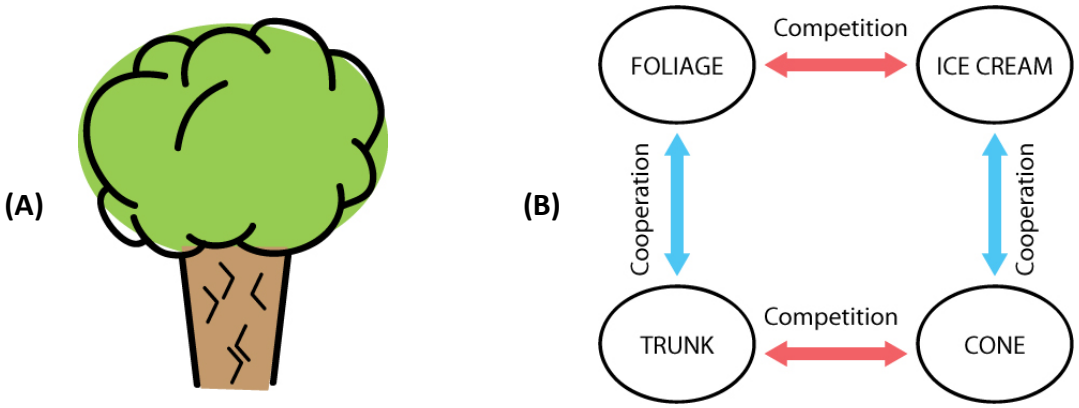


Figure 2.1-1: An ambiguous figure which can be interpreted either as an ice cream cone or a tree (A) with a schema assemblage formed to yield an interpretation (B) – perceptual schemas for ice-cream and cone cooperate, as do those for foliage and tree-trunk (blue arrows) while different schemas compete to interpret a given region (red arrows) (adapted from Figure 2.2.2 and Figure 2.2.3 of Michael A. Arbib, 1989).

The schema assemblage is an emerging pattern of a schema network through extensive processes of *competition* and *cooperation* among schema instances with various confidence levels (the competition and cooperation paradigm), which may invoke schemas beyond those initially associated with the scene. Cooperation occurs in 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. Competition occurs when there is conflict between schemas interpreting a particular region of a scene. For example, in Figure 2.1-1, the schema for "foliage" gets a boost for interpreting the region just above a region already interpreted as a tree "trunk", and vice versa, while the confidence level of the foliage schema may get reduced by the existing schema for "ice cream". Thus, a schema instance may initially become more active (i.e. assigned a higher confidence level) if it is consistent with more features of a region which it is competing to interpret. Cooperation then 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. For a scene perception task, the solution would be the interpretation of the scene where successful instances of perceptual schemas become part of the current representation in working memory.

2.2. VISIONS System

Within the framework of schema-based visual perception, we focus on the way in which perceptual schemas are associated with a visual scene to yield a semantic representation that can be used as the basis for generating a verbal description of the scene. More specifically, 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, the thesis builds on the language system's use of the result of visual processing – we do not offer new contributions to the schema-based study of vision itself.

An early example of schema-based interpretation for visual scene analysis is the VISIONS system (Draper, Collins, Brolio, Hanson, & Riseman, 1989), which deploys a set of perceptual schemas to label objects in a static visual scene. Although it has been several decades since the introduction of the VISIONS and there have been more advanced models of scene analysis and recognition (e.g. Li, Socher, & Fei-Fei, 2009; Sudderth, Torralba, Freeman, & Willsky, 2005), it still remains relevant to cite such "ancient" work because 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. Especially, 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 is particularly important when active "top-down" coupling from the language system back to the vision system is necessary – e.g. answering a question, such as "*what is it that John is holding?*", can be seen as identifying a missing slot in a scene schema.

However, note that in the VISIONS system, the ability to recognize a visual scene was limited in a sense that the general nature of the scene (e.g. a suburban scene with houses, trees, lawn, etc.) is prespecified, and only those schemas are deployed which are relevant to recognizing this kind of scene. Moreover, the current work extends the coverage of the original VISIONS system by incorporating the interpretation of events extended in time, especially the crucial addition of actions.

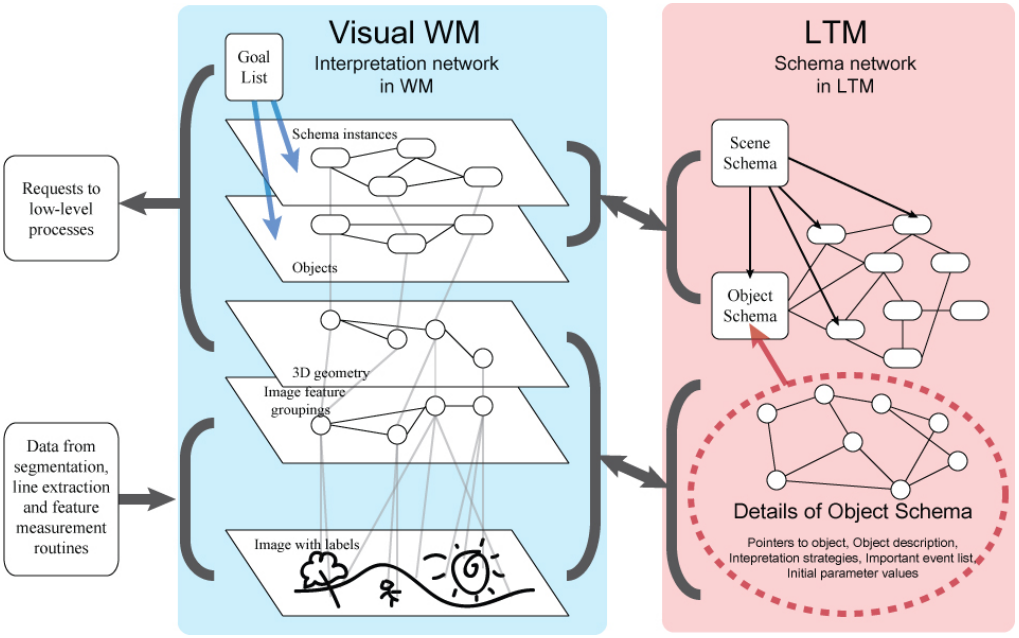


Figure 2.2-1: The Visual Working Memory (WM) of VISIONS interprets the current scene by a network of parameterized instances of schemas from Long Term Memory (LTM). These schema instances are linked to the visual world via the intermediate database that offers an updatable analysis of the division of the world into regions that are candidates for interpretation as agents and objects, possibly in relation with each other.

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 that may drive low level processes to re-segment the previously processed regions of the scene. Then the VISIONS system 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 long-term memory (LTM), while the state of interpretation of the particular scene unfolds in working memory (WM) as a network of schema instances. These schema instances are associated with specific portions of the image to represent aspects of the scene.

As specified according to schema theory, schema instances may compete and cooperate to determine which ones enter into the equilibrium schema analysis of a visual scene (Figure 2.1-1). Each schema instance in WM has an assigned confidence level which changes on the basis of interactions with other units in WM. Moreover, once several schema instances are active and make a coherent cooperative network, they may instantiate other schemas in a “hypothesis-driven” way – e.g. recognizing what appears to be a roof will activate an instance of the house schema that will bias the system to seek a wall in the region below that of the putative roof. Schemas with conflicting interpretations will compete for dominance over a certain region over the scene, and the losers will eventually be eliminated when their confidence level drops below a certain

threshold level. The system iterates the process of adjusting the activity level of schemas linked to the image through cooperation and competition until a coherent interpretation of (parts of) the scene is obtained.

2.3. SemRep: Semantic Representation for Visual Scenes

Although the coverage of the present work is to propose a computational model which mainly addresses the description of a given scene, the range of speech acts happening with the perception of a visual scene is vast: we can ask questions on specific aspects of the scene, make up a story, or draw attention to a certain object, to name a few.

During such processes, it is obvious that the vision system should be tightly coordinated with the language system regardless of the type of process – whether it is production or comprehension of speech. Such processes require a type of representation compact enough to be shared between the two systems via high-level cognitive processes but with enough details to be readily described to and reconstructed from a string of words.

Moreover, such a representation needs to be dynamic, even for a static scene, in its nature. The representation may be changed frequently by updates from the vision system if it perceives different scene objects, or a new object has appeared within the view. Also, the language system may send requests to the vision system for details of the scene that might have been overlooked previously (e.g. *what is it that John is holding?*).

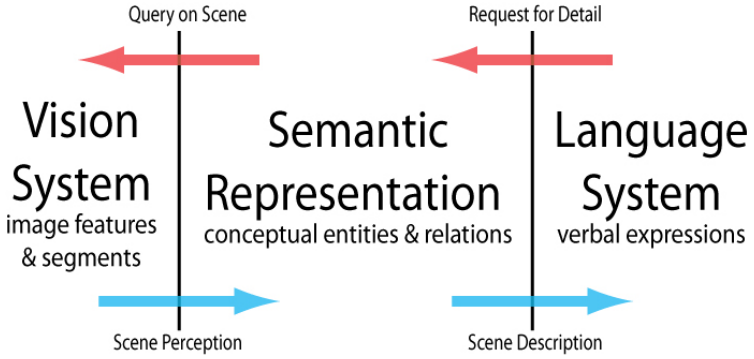


Figure 2.3-1: An illustration of the two-way interaction between the vision and language system. The role of the intermediate semantic representation is emphasized. It serves as a "dynamic bridge" between the two systems through which updates and requests are frequently exchanged. The upper (leftward) arrows indicate the requests from the language side asking for more information in the cases such as when a question is given (e.g. *who's kicking the ball?*), or when an underspecified object is about to be described. The bottom (rightward) arrows indicate the flow of such information delivered from the vision side to the language side for the production of description.

Our first contribution, going beyond the VISIONS framework, 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. For an initial effort, we introduced *SemRep* (the abbreviation of ***Sem***antic ***Rep***resentation) (Michael A. Arbib & Lee, 2007, 2008) as an encapsulation of what is in visual working memory which is organized into a form that is readily transferable for verbal expression. SemRep is basically defined as a hierarchical graph-like representation of a visual scene, whether static or dynamical over time (i.e. an episode). The semantics of an entity captured by perceptual schemas is reduced to a node or edge

to which is attached a concept while the semantics of a scene, also captured by perceptual schemas, is represented by adding the connectivity between those components. Thus, SemRep is an abstraction from the schema assemblages generated by the VISIONS system, which is explicitly designed to link the semantics of sentences to the representation of visual scenes.

Since a SemRep graph is regarded as an abstract representation of what is being perceived by vision, it only represents the semantics of "some" (not all) of the cognitively salient elements of the scene. SemRep may be viewed as an "interpretation" of the scene, rather than a "snapshot", as it is unlikely to capture all the subtle details of objects and events present in the scene. Only a set of perceptual schemas relevant to the current plans and goals for interaction with the environment are instantiated into a schema assemblage, and among those, SemRep abstracts out only a few "cognitively important" events, objects, and details. Even for the same scene, therefore, SemRep may result in a different graph at each moment, by capturing different aspects of the scene according to the given goals or the history of attention (e.g. from the event described in Figure 2.3-3, one might focus on the woman's hitting the man whereas the other focuses on her prettiness and the gaudy color of her dress).

A similar idea has been applied to the vision system proposed by Navalpakkam and Itti (2005). They used a particular topographic map, the Task Relevance Map (TRM), which highlights locations depending on their task-relevancy. The TRM acts as a top-down mask or a filter applied to bottom-up activation such that the system's sensitivity to perceptual salience of a location is raised or reduced according to the possibility for task-relevant targets to be found within that location. Although both SemRep and TRM share a common ground where only important components of a scene are highlighted and both can be used as a topological representation of a scene, the graph structure of SemRep allows a number of advantages over the simple overlay that the TRM provides.

First of all, SemRep can extend beyond what appears in the current scene. The topology of SemRep – i.e. the arrangement of conceptual entities and their connections – need not follow that of a scene. A description of a man without an arm, for example, can be represented as a node for the man and the node for the missing arm connected by an edge denoting the relationship as "missing". In fact, this representation does not exactly match an actual object setting since it "includes" a node for an arm that is missing in the actual image.

Moreover, SemRep can represent an event (or even multiple events) that happens over a certain time duration – i.e. a dynamic scene. The TRM, which is limited to a single depiction of a scene, cannot represent such an event. Again, the structure of SemRep does not have to follow the actual changes of the event, but it may contain only "conceptually significant" changes. 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. Note that this is a much more abstract level of description than that of a sensorimotor representation, such as the one captured by perceptual schemas, where continual tracking of task-related parameters is required.

Lastly, SemRep can be extended to represent an event happening in a 3-D space. Since SemRep is an abstract graph representation, a node can be associated with any arbitrary point in the space, representing an object set in a 3-D scene.

In sum, a prime motivation is to ensure that this representation be usable to produce sentences that describe a scene of

various situations, allowing SemRep to bridge between vision and language. Moreover, as emphasized by comparing with the TRM, the use of SemRep shares basic principles with other general approaches of semantic representation based on graphical structures, such as a *semantic network* (Sowa, 2006). Thus, we are confident of its extensibility to other meanings, especially since it includes actions and events extended in time.

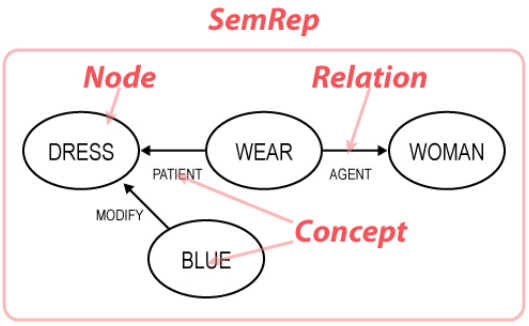


Figure 2.3-2: A schematic view of a typical SemRep. A SemRep is basically a graph-like structure that represents the semantics of a perceived scene.

As briefly mentioned earlier, a SemRep for a scene consists of a graph, which is basically a set of nodes and edges. Nodes are linked to cognitively significant entities and their regions of the scene, and their arrangements are configured by edges between them. As illustrated in Figure 2.3-2, we provide the definition of SemRep as follows:

I. A *SemRep* is a set of nodes and edges and their associated concepts, which are imposed on a scene as an abstraction from the schema assemblages that are resulted through perception of the scene.

II. A *node* is associated with a "cognitive entity" of a scene (e.g. object, agent, action, or attribute), and it is spatially "anchored" to the region occupied by the entity (see Section 2.5 for more detail).

III. A link, or an edge, between nodes specifies a *relation*. A relation specifies the relationship (e.g. spatial, possessive, componential, attributive, thematic, physical, or conceptual relationships, etc.) between cognitive entities.

IV. A *concept* is what describes the interpreted meaning of a cognitive entity (shown as a node) or a relationship between cognitive entities (shown as a relation). It is an abstraction of the semantic and syntactic knowledge derived from perceptual schemas with enough (but not more than enough) information to be translated into a verbal expression (see Section 2.4 for more detail).

V. A SemRep may be organized into a number of substructures, forming a *hierarchy* (see Section 2.7 for more detail). Each substructure represents a particular subevent of the scene captured within the SemRep while the conceptual hierarchy of substructures (e.g. "being a sub-part of", or "being less significant than", etc.) is reflected within their organization (not covered in the current work).

VI. A node and a relation may be given a *significance value* which expresses the cognitive importance of a particular aspect of the scene. Factors such as goal-relatedness, familiarity, or perceptual salience may contribute to a significance value (not covered in the current work).

Consider the specific scene and SemRep shown in Figure 2.3-3. The visual system may initially recognize a variety of

aspects of the scene centered around the central figures of the man and woman, while ignoring other aspects of the scene. This analysis may combine activation of a number of schema instances together with activity in the intermediate database that could be used to support further schema analysis, but has not yet done so. SemRep then abstracts from this pattern of schema activation a set of nodes and relations which constitute one possible semantic structure for the current scene – a given scene may be perceived in many different ways.

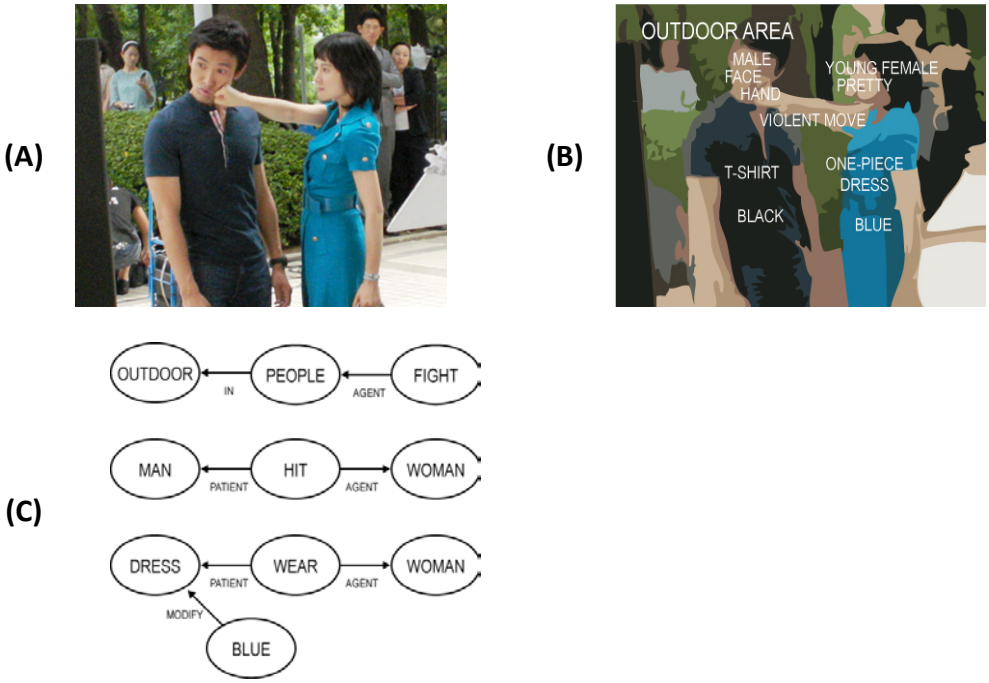


Figure 2.3-3: An example of a visual scene and SemReps possibly generated from the scene. (A) is a picture of a woman hitting a man (original image from: "*Invisible Man Choi Jang Soo*", Korean Broadcasting System), whose regions are segmented and labeled with perceptual schemas (whose parameterization is not shown) through a stage like the Visual Working Memory (Figure 2.2-1) of VISIONS (B). Note that although we use words to label these regions, this is for our convenience – words are not parts of these schemas themselves. (C) illustrates three example SemRep graphs that could be possibly generated for the events illustrated in the scene, depending on the cognitive prominence of each event – one for the event of the woman hitting the man, one for the woman's wearing a blue dress, and the other for the man and woman's fight happening in an outdoor area. Again, the words on the nodes are labels of convenience for yet-to-be-verbalized concepts. Theses SemReps might yield such sentences as "*a woman hits a man*", "*a woman is wearing a blue dress*", or "*people are fighting outside*".

Here, agents and an action are represented as nodes and both nodes and relations are labeled with concepts (in capital letters) – i.e. the recognized or interpreted meanings are attached to the node for that instance of the object, and the semantics of an action are attached to an action node. We also use nodes to represent attributes since they also capture independent concepts from those of objects that they decorate (but they may occupy the same regions). A node for action generally has relations with other multiple nodes, each playing a different thematic role of the action event. For example, except for a node for the action itself, a transitive action involves two other nodes that are the agent and the patient of the action, respectively. A

more complex form of action, such as expressed in English using the ditransitive case, may have more relations with entities of different thematic roles.

When the nodes and relations of a SemRep are translated into words by the language system, the attached concepts help resolve the lexical selection process by providing semantic clues. However, one should not confuse the concepts of the SemRep graph with specific word labels since concepts are more abstract descriptors of semantics which eventually allow 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. Moreover, the same SemRep can be the basis for description in any language once the appropriate grammar and lexicon are deployed.

Some of these processes may be directly perceptual, possibly generated immediately by the visual system, while others may be more inferential, possibly drawn from the involvement of further world knowledge through the propagation of activation through the semantic network.

2.4. Semantico-syntactic Features

Conceptual representations, which are associated with schemas and their abstraction as captured within the concept of the SemRep, are proposed to be neurally grounded in sensory and motor systems of different modalities (see Section 3.1 for the detailed account on the neurophysiological establishment for semantics and concepts). Especially, the perceptual states across modalities (e.g. visual, tactile, or olfactory, etc.) and their integrations that are shaped through sensorimotor experiences are captured and delivered by perceptual schemas, and the role of these perceptual schemas within the action-oriented perception framework is to provide relevant perceptual states to motor schemas that control the agent's embodied interaction with the world. Therefore, the concepts encoded within SemRep are not mere symbols that are detached from sensorimotor representations but rather "perceptually grounded" in the sense that their meanings are associated with specific perceptual experiences and the coupled motor actions, highlighting the "embodiedness" of such representation.

Glenberg (1997) proposed a very similar view to our action-oriented perception framework, arguing that memory and conceptualization work in the service of perception and action – being guided by memory, conceptualization is the encoding of patterns of possible physical interaction with a three-dimensional world. His view is essentially "embodied" because it is implied that how we perceive and conceive of the world is determined by the types of bodies we have. These so called "embodied" theories are recently getting more popularity as supporting empirical evidence is increasingly reported (Kan, Barsalou, Solomon, Minor, & Thompson-Schill, 2003; Pecher, Zeelenberg, & Barsalou, 2003). Among those theories, for example, the perceptual symbols approach (Barsalou, 1999; Barsalou, Simmons, Barbey, & Wilson, 2003) emphasized the use of sensorimotor representations and embodied experiences to ground "perceptual symbols" in the human cognitive system. Association areas in the brain that capture bottom-up patterns of activation in sensorimotor areas during perceptual experience partially reactivate sensorimotor areas to implement perceptual symbols in a top-down manner. This "re-enactments" or "simulation" of states in modality-specific systems, which is claimed to implement basic conceptual processing, is also emphasized as the key component of the system.

However, although we generally agree with the importance of sensorimotor experience in grounding our conceptual representation, we do not totally agree with the strong version of the embodied view. In his critique of Gallese and Lakoff's (2005) claim, Arbib (2008) argued that semantics and grammar have their roots in specific sensorimotor experience but have developed (both historically and ontogenetically) through layer upon layer of abstraction to handle concepts which are not embodied save through their history, thus allowing even descriptions of counterfactual events like "*Pegasus is flying through the winds of Jupiter*". He further argued that although some conceptual distinctions indeed arise from motor and perceptual discontinuities, others are still imposed top-down by some type of "symbolic overlay", such as language, serving to anchor conventionalized distinctions – e.g. the base-level category WHALE does not entirely depend on embodiment or our human existence, but rather it depends on the knowledge of biological science such that we place WHALE under the MAMMAL hierarchy rather than under FISH. Similarly, Mahon and Caramazza (2008) also pointed out the limited scope of the embodied cognition framework, suggesting that an embodied theory of cognition would have to admit "disembodied" cognitive processes in order to account for the representation of abstract concepts, such as JUSTICE, ENTROPY, BEAUTY or PATIENCE, whose "meaning" corresponds to no sensory or motor information in any reliable or direct way.

Thus, we share our position for the conceptual representation with the studies mentioned earlier such that there is a level of conceptual representation that is abstract and symbolic enough to be "compositional" (Michael A. Arbib, 2008; Mahon & Caramazza, 2008) yet is complemented, or even enhanced (Fischer & Zwaan, 2008), by the information represented in the sensory and motor systems. Recently, Negri and colleagues (2007) reported a case of stroke patients whose ability to recognize actions (including pantomimes) and objects dissociates from the ability to use those same objects, rejecting the strong form of the embodied cognition hypothesis. Furthermore, neurophysiological data supports the existence of high-order association areas (Binder & Desai, 2011; H. Damasio, Grabowski, Tranel, Hichwa, & Damasio, 1996; H. Damasio, Tranel, Grabowski, Adolphs, & Damasio, 2004; A. Martin & Chao, 2001; Murray & Richmond, 2001; Rogers et al., 2004; Vargha-Khadem, Gadian, & Mishkin, 2001), implying that the cognitive system might employ abstract and even symbolic representations to some extent.

The embodied theories were originally proposed – while unfortunately ignoring the action-oriented approach already in place from schema theory – in reaction to the "disembodied" theories (e.g. Fodor, 1998; Pylyshyn, 1984; Smith & Medin, 1981), and the debate between the disembodied and embodied approaches has recently heated up (Gallese & Lakoff, 2005; Mahon & Caramazza, 2005). The disembodied theories assume that conceptual representations are amodal and symbolic and they operate according to different principles than representations in modality-specific systems, and that knowledge resides in a modular semantic system separate from modality-specific systems for perception, action and emotion. On the other hand, the embodied theories, especially in their strong form, propose that the sensorimotor system satisfies all principal criteria for characterizing both sensorimotor and more abstract concepts (Gallese & Lakoff, 2005). The debate between the two arguments is centered around a dichotomy between states in modality-specific systems and redescriptions of these states in amodal representational languages to represent knowledge (Barsalou, et al., 2003).

However, one should note that our concerns of the present work differ from the debate; our approach taken here does not focus on the judgment of whether "concept" is an abstract and symbolic representation detached from sensory and motor systems. Rather, we emphasize multi-modal and context-dependent integration across sensory and motor systems and the

resulting assemblages of perceptual and motor schemas over the single concept. The action-oriented perception framework claims that the type of information available within the schema assemblage being formed in an agent's WM is dependent on the particular type and course of actions that are deployed to meet the goals of the interaction that the agent is to perform. Therefore, depending on context, the perceived concept of an "apple" may mean many different things – e.g. an apple as the fruit to be eaten, an apple to be picked up, or even an apple to be verbally described. The implication is that among all the possible APPLE concepts (if these can be counted), only a few concepts contribute to the meaning of the apple at a certain moment.

Given the role of SemRep as the mediating representation between the vision and language systems, we can claim that the semantic information encoded within a SemRep comprise mainly the particular type of information required for the action of verbal description. Thus, here we propose that a concept mostly encodes the "semantico-syntactic" knowledge of an entity; the concept of a perceived entity represents an encapsulation of the semantic and syntactic knowledge derived from perceptual schemas, which is compact but detailed enough to allow application of lexical constituents and sentential structures and eventual translation into a verbal expression. Although the specific type of features encoded within a concept may vary depending on the language to be spoken, a concept generally conveys semantics-oriented features such as animacy, categorical knowledge (e.g. TIGER is-a ANIMAL), and thematic role (e.g. AGENT does ACTION to PATIENT), as well as syntax-oriented (yet still semantically grounded) features such as gender, person, number, tense, and definiteness.

However, one should note that application of SemRep is not limited to scene perception for description but also could be extended to more general cases of scene perception. As the type of information contained within a concept varies according to the task, the semantico-syntactic knowledge is chosen only because the particular type of action for the task is verbal description. Thus, in an extreme case where no linguistic action is required, such as free-viewing of a scene or searching for a particular object, concepts within a SemRep may only contain semantic features rather than any syntactic features. Moreover, even during the task of scene description, SemRep may contain more perceptually-detailed semantic features when necessary – e.g. when the identity of an object needs to be further resolved.

The concept as we propose here for verbal action appears to be similar to lexical representations like *lemma* (Bock & Levelt, 1994; Roelofs, 1992) or *prominence* (Bornkessel-Schlesewsky & Schlesewsky, 2008; Bornkessel & Schlesewsky, 2006). As opposed to the lemma, however, the concept encodes more semantically defined features, such as categorical knowledge. While the lemma level specifies syntactic properties, such as grammatical class (noun, verb, etc.), gender, and auxiliary type (be or have), semantic features are encoded in an independent level (the conceptual level) dedicated for lexical "concepts". Moreover, prominence has been proposed to contain "rigorous" knowledge of syntactic structure, such as morphological case marking or constituent order, whereas the concept within our framework does not contain such syntactic information (although it contains some of syntactic features). As will specified in Section 4.3, such information is handled within the level of construction, not in the level of semantic constituent or object concept. Most crucially, the semantico-syntactic knowledge encoded in a concept is proposed to be "language-specific" as opposed to prominence whose development is claimed to be cross-linguistically motivated.

Since the processing of the concept may frequently be augmented with more detailed conceptual knowledge or sensorimotor representations, and the concept may include more or less semantic information depending on the type of

language and the linguistic task, it is difficult to set a hard boundary as to how much semantic information is encoded within a concept. Nevertheless, a distinction needs to be made between the type of information represented within a concept and more generic conceptual knowledge like world knowledge. Compared to the latter, the former is “shallower” in that it comes with just an enough amount of information to be translated into a verbal expression or to be used in linguistic processes. On the other hand, the latter is “deeper” in the sense that it provides more ample knowledge of the world and sensorimotor information, thus allowing us to draw an analogy such that the latter representations are like entries with detailed descriptions in an “encyclopedia” while the former representations resemble more concise and language-oriented entries in a “dictionary”.

However, we do not propose that those two types of representation either exist in a distinctive manner or should be treated separately; the former type of information may be constantly complemented by, or may even incorporate, the latter type of information when required as specific sensory and motor representations provide rich contextual information during linguistic processing of a concept. Moreover, within the framework of schema theory, both type of information is represented in terms of schema assemblages (mostly of perceptual schemas in the level of the SemRep) that are activated from a schema network, and “further activation” of the schema network enriches the processing when necessary.

2.5. Indexing Entities

Even though the initial work on the visual pathways can be exemplified as specifying “what” object is “where” (Mishkin & Ungerleider, 1982), Milner and Goodale (1995) claimed that it is more appropriate to call the dorsal pathway the “how” pathway because the distinction of the visual pathways is not between subdomains of perception (“where” and “what”), but between perception (“what”) and the guidance of action (“how”). They argued that the purpose of the dorsal pathway is to visually guide actions by providing many properties needed to determine how to interact with an object, with location (“where”) being only one of those properties. This idea was initially proposed by Goodale and colleagues (1991) who emphasized the distinction between the neural substrates of visual perception and those of visual control of actions.

Milner and Goodale’s claim has been further supported by lesion studies on the patients with impairments in these pathways. There were reported cases where a patient (DF) with a ventral lesion was able to carry out a variety of object manipulations even though unable to verbally report or pantomime the object parameters used to guide these actions (Goodale, Jakobson, & Keillor, 1994; Goodale & Milner, 1992). For example, when she was asked to pick up objects with various sizes or orientations, she could preshape the hand accordingly, but she could not indicate the size or the orientation of the objects verbally or manually. Conversely, another patient (AT) with a lesion in the dorsal pathway exhibited the opposite deficit (Jeannerod, Decety, & Michel, 1994). While she was able to pantomime the size of a cylinder, she could not preshape appropriately when asked to grasp it.

Based on the ventral and dorsal dissociation described above, Fagg and Arbib (1998) proposed that the dorsal pathway provides *affordances* of a visually perceived object, noting that the purpose of the dorsal pathway is to provide parameters for how to interact with the object while the ventral pathway provides the specific context presented by the object. Affordances (Gibson, 1986) are basically parameters for motor interaction that are signaled by sensory cues from vision or other modalities (Greeno, 1994) without invocation of high-level object recognition processes. In fact, in a primate study, Murata and colleagues (2000) reported that during grasping tasks on 3D objects, information like affordances, such as the shape, size,

or orientation of the objects, were encoded in a portion of neurons in the anterior intraparietal area (AIP).

Therefore, in terms of schema-based visual perception, we can divide the types of information conveyed through perceptual schemas broadly in two categories: the "ventral" parameters for recognizing and identifying objects, and the "dorsal" parameters (or affordances) for guiding motor actions on objects. Within the current framework where perceptual schemas are abstracted into a SemRep, the concepts are thought to encode the type of information corresponding to the ventral parameters since they represent the semantic properties of the associated entities. But where are the dorsal parameters encoded within a SemRep? Since the SemRep is currently proposed as a semantic representation for verbal description of a scene and the type of information required in linguistic processes is mainly "conceptual", we highlight concepts and their associations with perceptual schemas while emphasizing semantico-syntactic features that are encoded within concepts (Section 2.4). Thus, it seems unlikely that a SemRep encodes any "dorsal-like" parameters that are directly relevant to motoric actions for verbal description – e.g. specific phonetic information or articulatory commands for naming elements in a scene.

However, the type of task addressed in the current framework is not merely linguistic but rather "visuo-linguistic", and this implies that not only the language system but also the vision system is involved in the process. The visual representation that we form is far from complete and it continuously needs to be updated when more detail is required (Ballard, Hayhoe, & Pelz, 1995; Hayhoe, Bensinger, & Ballard, 1998; O'Regan, 1992; Spivey, Richardson, & Fitneva, 2004) – the visual world paradigm. During the perception task, the eyes need to frequently "peek back" to the entities under recent attention to extract more information, making some type of object locating mechanism crucial. This type of mechanism is even more important when the task involves a relatively long discourse where a number of different entities need to be referred to frequently (e.g. "*there is a boy and a girl…*", "*he is …*", "*but the girl is …*", etc.), or when a dynamic scene, such as a videoclip, is being described, where objects are constantly moving, occluded or reappearing.

Therefore, there is necessity for locating and tracking entities, and we propose that SemRep provides not only the conceptual knowledge of entities but also some type of "location coordinates" of those entities (moving or nonmoving) – if you recall that each node in SemRep is "spatially anchored" (Section 2.3). We may view such coordinates as a type of "affordances for visual actions". Similar to affordances for grasping actions, these affordances provide a set of parameters, such as spatial coordinates with respect to the perceived entities, for oculomotor actions or attention shifts. Once an entity is recognized and represented as a node in a SemRep, the entity is bound with the node through a type of indexing mechanism so that subsequent processes can reliably access the entity for more information, establishing a full association between the node and the entity (i.e. dorsally and ventrally). This type of indexing mechanism can be used for such cases as gaze fixation, directing attention, or object pursuing. All of those cases are not necessarily limited to a scene description task, implying that SemRep can be used for a more generic framework of action-oriented perception (e.g. perceiving a scene for a grasping action).

A number of studies also emphasized the importance of such indexing apparatus, arguing for the visual world paradigm. These studies claimed that during visual perception, spatial indices are allocated to the regions of the perceived scene in order to aid in sorting and separating the events that took place in them, with eye movements being the accessing method (Altmann, 2004; Richardson & Spivey, 2000; Spivey, et al., 2004). The perceived scene is transformed into an internal image

constructed in a "visual buffer" (i.e. visual WM) where the access to the certain parts and aspects of the scene is done by shifting attention to those locations (Kosslyn, 1994). It appears that such an internal representation is in principle compatible with the SemRep.

As briefly mentioned earlier on the dorsal-ventral distinction, the parameters for guiding visual actions (e.g. shifting attention, saccades, etc.) are more "dorsal" while the parameters that provide information on the semantics and concepts of perceived objects are considered to be more "ventral". These two sets of parameters are supported through the two distinctive streams of visual perception (see Section 3.2 for more on the visual perception network proposed in the current framework). Especially, the dorsal parameters (or visual affordances) are processed by the circuitry within the dorsal stream, such as the posterior parietal cortex (PPC) or the frontal eye field (FEF) – given the involvement of these regions in controlling attention shifts and eye movements, it seems reasonable to associate these regions with the indexing mechanism.

Furthermore, we would like to emphasize that the indexing mechanism subserves orienting attention, rather than occulomotor actions *per se*, with those actions being produced as the outcome of attention shifts. According to the results from the multiple object tracking (MOT) experiments reported by Pylyshyn and Storm (1988), most subjects were able to continuously keep track of as many as 4 or 5 targets over several seconds, and they seemed to have used a type of visual indexes that are assigned to the items being tracked (FINSTs; see Pylyshyn, 2001 for more details). Although the detailed nature of this indexing mechanism is still unclear, it seems highly unlikely that eye movements alone can explain the tracking capability of multiple targets (Cavanagh & Alvarez, 2005), strongly suggesting that the indexing mechanism does not directly guide eye movements, but rather it is more related to attentional processes. In fact, some evidence indirectly supports this view – e.g. attentional pursuit (indexing or tracking) is substantially faster than successive saccades between objects, suggesting different supporting machineries (Horowitz, Holcombe, Wolfe, Arsenio, & DiMase, 2004).

Moreover, evidence suggests that areas in the dorsal pathway, especially posterior parietal regions, are implicated in controlling attention shifts (Corbetta, Kincade, Ollinger, McAvoy, & Shulman, 2000), and some studies further associated these regions with "motor attention (i.e. directing attention to manual or more general movements)" (Kawashima et al., 1996; Rushworth, Johansen-Berg, Göbel, & Devlin, 2003). Especially, results from monkey studies strongly indicate that the spatial locations represented by neurons in the lateral intraparietal area (LIP) of the PPC reflect the locations of attentional focus (e.g. cognitively salient locations, where the animal's intention is directed, etc.) rather than simple saccade coordinates (Bisley & Goldberg, 2003; Gottlieb, Kusunoki, & Goldberg, 1998; Snyder, Batista, & Andersen, 1997; Williams, Elfar, Eskandar, Toth, & Assad, 2003).

Further evidence comes from lesion studies on patients with "simultanagnosia" who have disruption in the dorsal pathway (bilateral parietal damage that sometimes extends to occipital regions) that is associated with the inability to perceive simultaneous events or objects in their visual field – they cannot see more than one object at a time while their perception of individual objects remains intact, resulting in failure to grasp the overall meaning of the image (Farah, 1990). Various neurophysiological and behavioral studies have suggested that the deficit mainly originates from failure in attention control and the related mechanisms. More specifically, it is caused by the inability to establish or maintain the linkages between perceived objects and the appropriate location (Coslett & Saffran, 1991), the inability to keep track of spatial locations of perceived objects (Dehaene & Cohen, 1994), the inability to inhibit attentional bias toward irrelevant (but

salient) stimuli (Karnath, Ferber, Rorden, & Driver, 2000), or the inability to bind visual feature information (e.g. binding color and form) into a coherent, perceptual unit (McCrea, Buxbaum, & Coslett, 2006).

Therefore, the SemRep not only represents the semantics and concepts of the perceived entities in a scene but is grounded in spatial indexes relevant for guiding attention to those entities. In that sense, the locational information encoded in the SemRep can be regarded as a type of affordances as they are distinct from mere spatial coordinates – i.e. more of "how" than just "where". Moreover, emphasis has been given to attention orienting and control, rather than to physical eye movements, as the process that is guided by such indexing mechanism.

Moreover, as evidenced by neurophysiological and behavioral studies so far, the type of spatial information for the indexing mechanism and guiding attention is grounded in the dorsal system. This necessitates a distinction from the type of spatial information subserved by the ventral system, which is generally represented as spatial configurations between entities in the SemRep (i.e. relations and associated concepts). Kosslyn (1987) has proposed that the visual system uses two types of spatial relations – "categorical" representations capture general properties of the spatial structure of a visual stimulus (e.g. "*this line is 'above' the two dots*"), without defining the exact metric properties, while "coordinate" representations specify precise spatial locations of objects or parts in terms of metric units (e.g. "*these two dots are 1.6 cm apart*" or "*this line can be fit in between the two dots*"), and this claim is supported by a number of subsequent studies (Jacobs & Kosslyn, 1994; Jager & Postma, 2003). Although the specific nature of the claim is not certain yet, we see this claim emphasizes the different between the two types of spatial information delivered within a SemRep – one is more conceptual (ventral) and may be view-point invariant (i.e. allo-centric) while the other is more motor-oriented (dorsal) as it provides more precise metric parameters (i.e. object- or ego-centric).

In a series of non-human primate studies with his colleagues, Ma (2011; 2004; 2003) suggested a neurophysiological distinction between these two types of spatial information, claiming that the dorsolateral prefrontal cortex (DLPFC) and the medial prefrontal cortex encode the ego-centric spatial frame which provides a frame of reference for attention deployment whereas the hippocampus is the region related to the allo-centric spatial frame apparently for reflecting geometrical relationships between environmental cues to identify spatial location.

This neural separation in processing two types of spatial information is further supported by other studies. For example, it has been suggested that areas around the LIP encode locations and objects of interest in several ego-centric reference frames (Colby, Duhamel, & Goldberg, 1995; Colby & Goldberg, 1999; Gottlieb, et al., 1998; Mullette-Gillman, Cohen, & Groh, 2009), which are claimed to be a hybrid of different reference frames that are dynamically transformed (e.g. between receptor surfaces and effectors) according to the type of the required action. On the other hand, the hippocampus has been suggested to support viewpoint manipulation (King, Burgess, Hartley, Vargha-Khadem, & O'Keefe, 2002) and building relational structure (O'Keefe, 1999; Pierrot-Deseilligny, Müri, Rivaud-Pechoux, Gaymard, & Ploner, 2002) in spatial memory, and the parahippocampal place area (PPA) has been suggested to be a locus in processing spatial layouts (Epstein, Graham, & Downing, 2003; Epstein, Harris, Stanley, & Kanwisher, 1999; Epstein & Kanwisher, 1998). Moreover, the perirhinal cortex was claimed to play a primary role in object identification, binding the different views of an object with its attributes into a reliable representation, and associating objects with other objects (Murray & Richmond, 2001).

Therefore, the spatial information encoded within a SemRep may be divided into two separate types: the "ventral" type,

which provides, sometimes viewpoint-invariant, spatial context, and the "dorsal" type, which provides relatively direct means to locate and access entities. Despite this distinction in their nature, both types of coordinate systems are necessary in building a scene representation as one provides a constant spatial frame where object positions are contextually defined relative to the other objects while the other helps locate and direct attention to objects for visual perception. As addressed in more detail in Section 3.4, this particular feature of SemRep signifies the "procedural" aspects relevant to the continued availability of further (visual) updates, which plays an important role in the integrative processes of vision and language.

2.6. Hierarchical Scene Perception

In an experiment, Duncan (1984) showed subjects a brief display of two pairs of a box with a single line drawn through it. Both the box and the line varied in two dimensions: the box could be tall or short with a small gap on its left or the right side, and the line could be either dotted or dashed while leaning slightly to the left or the right. Interestingly, subjects were less accurate at reporting two properties from separate objects (e.g. the size of the box and the orientation of the line) than reporting two properties of a single object (e.g. the size of the box and the side of its gap), showing a same-object advantage. Based on this result, Duncan proposed that there are parallel preattentive processes that serve to segment the field into separate objects, followed by a process of focal attention that deals with only one object at a time. The implication of Duncan's proposal is twofold: (1) visual selection is an object-based (not space-based) serial mechanism, and (2) there are two separate stages in the processes of visual perception.

In this section, we first address Duncan's proposal and subsequently discuss its implications in the processes of visual attention and scene perception.

Visual attention is thought to be a selective mechanism concentrated to a specific area of the visual scene while its processing is performed in a serial fashion (Cave & Bichot, 1999; M.-S. Kim & Cave, 1995). The area covered by visual attention seems to be flexible (i.e. zoom-in and -out) with loss of resolution or efficiency for a larger size (zoom lens model; Eriksen & St. James, 1986) and tightly correlated with planning and performing saccades (Deubel & Schneider, 1996). As Duncan (1984) proposed, the area of visual attention is likely to be delineated in terms of objects rather than locations (Chen, 2003; Duncan, 1984; Nissen, 1985) although there has been a debate on this issue (Cave & Bichot, 1999). In fact, studies suggested that visual attention is directed not to just simple collections of features (Scholl, Pylyshyn, & Feldman, 2001) but to discrete, yet non-rigid objects (Yantis, 1992) that are not even necessarily fully identified (Kahneman, Treisman, & Gibbs, 1992).

Based on various findings on visual attention, however, Scholl (2001) argued that objects and locations should not be treated as mutually exclusive since attention may be object-based in some contexts, location-based in others. For instance, the units of selection seem to be complex enough to include spatiotemporal properties as well – targets can be tracked when they disappear behind an occluder or even when all objects disappear from view as in an eye blink (Horowitz, Birnkrant, Fencsik, Tran, & Wolfe, 2006; Scholl & Pylyshyn, 1999). Similarly, even 10-month-old infants were able to use spatiotemporal information to set up representations of distinct objects (Xu & Carey, 1996), and a group of objects moving in a common direction were reported to be treated as a single global object representation in an object tracking task (Suganuma & Yokosawa, 2006). Therefore, the "objecthood" of an attentional focus need not be specific to a particular property but it may

become a broader notion, which might have been shaped through the perception experiences of the observer (Pylyshyn, 2001).

Most evidence supporting the object-based view of attention came from the experimental results of a multiple object tracking (MOT) paradigm, in which subjects are required to track specific target objects among a number of distractor objects that are moving simultaneously (Pylyshyn & Storm, 1988). It has been reported that people are generally able to follow 4 or 5 targets by using some type of a preattentive indexing mechanism (see Section 2.5 for related accounts within the current framework of SemRep), and each tracked target can receive focal attention in a serial fashion when further processing is required (Pylyshyn, 2001).

Although emphasis has been given to the unit of attentional selection, the results of MOT experiments also imply separate processing stages in visual perception as Duncan (1984) proposed – the preattentive parallel processing stage for indexing multiple targets and the successive processing stage of serial focal attention.

Moreover, a number of studies based on a rapid serial visual presentation (RSVP) paradigm – stimuli such as letters, digits or pictures are presented successively at a single location at rates between 6 ~ 20 items per second – also suggested a similar separation of the process: at the first preattentive level, targets and nontargets are distinguished in a parallel fashion, and then at the second attentive level, only the selected target is stored and processed within a limited capacity system (i.e. visual WM) (Chun & Potter, 1995; Potter, 1976). The selection and identification of targets were reported to take 200 ~ 500ms as presentation of another target within that duration impaired detection performance, resulting in a phenomenon known as the attentional blink (AB; Raymond, Shapiro, & Arnell, 1992). Similarly, Awh and colleagues (2006) argued that the processes of selective perception are composed of multiple stages of processing including both early sensory (enhancement and inhibition of sensory features) and postperceptual processes (active maintenance of information in working memory) with attention acting as a gatekeeper to the later stage.

A line of studies on "subitizing" also suggest a distinction between the preattentive process and the successive attentional process. Subitizing refers to the process of effortless and rapid recognition of the number of items within the visual scene, when the number of items falls within a certain range, which is generally capped around 4 (Trick & Pylyshyn, 1994). Subitizing is contrasted to counting, which involves serial deployment of attention on each item. The experimental result on patients with simultanagnosia, who have deficits in attentional processes, indicated that subitizing and counting are supported by different mechanisms – patients showed relatively spared performance in subitizing with smaller sets of 1, 2, or sometimes 3 items while they are impaired with counting larger sets (Dehaene & Cohen, 1994). A study on a working memory task reported a similar dissociation where subjects' working memory task score was only associated with their performance on the attention-demanding counting portion of the enumeration task (Tuholski, Engle, & Baylis, 2001). The suggestion is that subitizing is supported by preattentive parallel processes, which might be of limited capacity (e.g. Lavie & Cox, 1997), while counting is performed by postperceptual attention processes in a serial manner.

Note that subitizing and the indexing mechanisms in MOT tasks are both claimed to be preattentive, with their maximum capacity around 4, and this might suggest a strong connection between those two processes. In fact, Pylyshyn (2001) emphasized the similarity and claimed that subitizing is supported by the indexing mechanism for MOT tasks – he proposed a special kind of direct connection to items in the visual field, which he named as FINSTs (FINgers of INSTantiation), for

such an indexing mechanism.

Rensink (2000a, 2000b) proposed a model of scene perception based on an architecture which basically consists of three components, one for processing volatile low-level features to form structures (proto-objects), one for non-attentional processes that provide a gist or a layout to guide attention, and the other one for attentional processes to build a coherent representation of a scene or an object. The key feature of his proposal is a notion of "coherence field", which is a dynamic form of representation where attention provides detailed, coherent descriptions of an object through the established "links" that are attached to properties of the object (e.g. low-level visual features, proto-objects, object parts, etc.). These links are intrinsically similar to the type of spatial anchors that ground the indexing mechanism (e.g. FINSTs) since they provide the more stable viewer- (or object-) centered coordinates from the ever-changing retinotopic coordinates, through which the captured low-level structures (proto-objects) can be managed. His architecture is also based on the idea of the preattentive-attentional dissociation discussed so far since it involves an initial extract of gist and subsequent refinement of detail with a coherence field acting as an intermediate representation bridging the two processing stages.

An interesting aspect of Rensink's proposal is that a coherence field forms a local "hierarchy" with two levels (object- and part-level) of description. The links are assigned to parts of an object being represented within a coherence field and attentional processes can traverse up and down through these links, thus establishing a whole-part hierarchy of the object. As Marr (1983) insisted, such whole-part hierarchy is a natural way to represent (visual) objects, and it has been exploited in several models of visual perception (e.g. Deco & Schürmann, 2000).

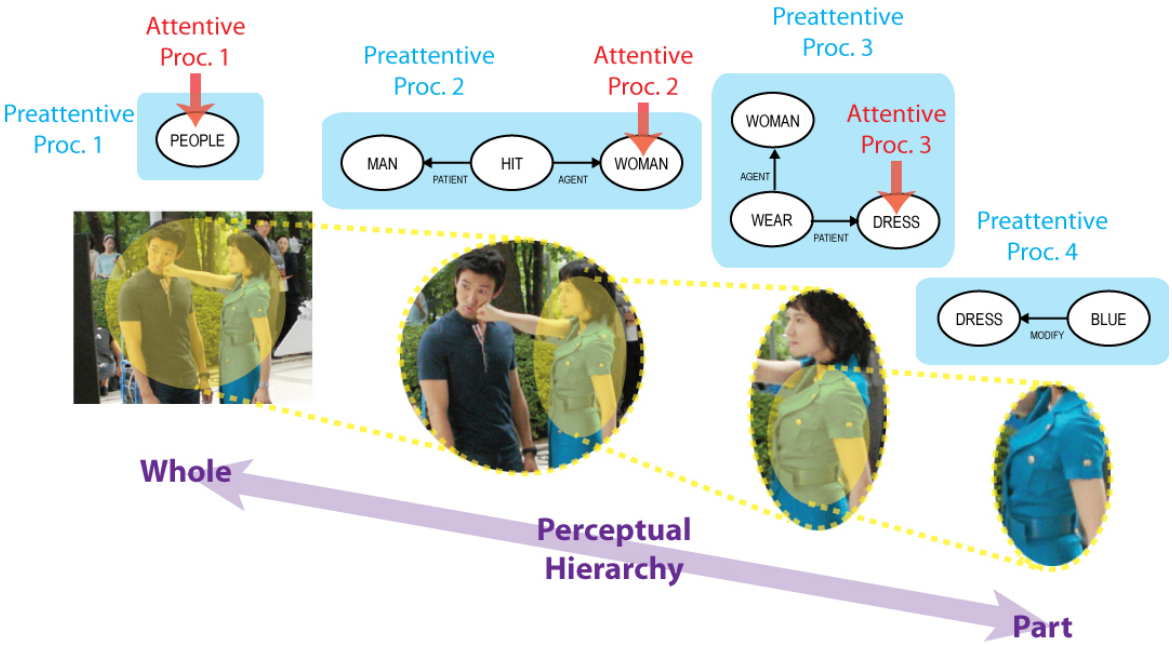


Figure 2.6-1: A schematic view of hierarchical perception process. As attentional focus moves up (zoom out) and down (zoom in) the hierarchy, the early nonattentive and the later attentive process are executed as a pair at each level, forming an iterative cycle through the whole-part hierarchy (four cycles shown). A gist or a layout (shown as a SemRep) is perceived at each level, guiding attentional focus.

In fact, the early-late dissociation in the process of visual perception, as addressed by the studies covered so far, is

intrinsically “iterative” – the early preattentive process (for “whole”) and the later attentive process (for “part”) form an iterative cycle of processing through the whole-part hierarchy (as illustrated in Figure 2.6-1). More specifically, visual perception of an object is done in such an iterative manner that the early parallel processes deliver the layout of the object based on the information of the object parts, which are presumably captured by the indexing mechanism, while each of these parts will be successively focused (zooming in) when further processing is required, and then the focused part will be again processed by the early parallel processes as a whole, and some of its parts in turn get thorough examination under attentional focus. Once a close inspection on a part is done, attentional focus may move up to the whole-object level (zooming out) possibly to focus on another aspect.

Hochstein and Ahissar (2002) proposed a similar framework (Reverse Hierarchy Theory) in which they defined two divided processes: the initial “vision at a glance”, which is automatic, wide-scale, and implicit bottom-up processing, and the later “vision with scrutiny”, which is more indirect conscious perceptual constructs by focused attention. They argued that these two processes cooperate during visual perception in such a way that the initial process builds a basic-category-level coherent percept by spreading attention and guessing at details, which is error-prone, while the later process explores details and resolves conflicting features, eventually building a subordinate-category-level perception. Similarly, Henderson and Hollingworth (1999) argued that initial fixations are controlled by global visual features or concept of a scene, but as viewing progresses, fixations are controlled by the visual and semantic properties of the local regions.

This dissociation of processing stages in the vision system seems to reflect how our cognitive system perceives and stores visual representations. Our cognitive system exploits the whole-part hierarchy of vision to reduce the workload for building and storing perceived visual representations while minimizing processing of unnecessary details. As the visual world paradigm (Ballard, et al., 1995; Hayhoe, et al., 1998) insists, the visual representation that we form is extremely scant with only minimal information carried in while the details are collected as they become necessary for the current task or goal.

“Change blindness” (for a review, see Simons & Rensink, 2005), which is defined as the failure of observers to detect large, sudden changes in a display, well exemplifies such “sparseness” of visual representations, and a number of studies suggested that information worth only a few (around 4 or 5) objects is stored in visual short-term memory at one time (Irwin & Zelinsky, 2002; Luck & Vogel, 1997). Especially, the preattentive processes presumably grounding for subitizing and the indexing mechanism was reported to lack detailed features of items, such as colors or shapes (Scholl, Pylyshyn, & Franconeri, 2004).

Thus, the vision system should be supported by processes with more focused attention in order to fully “scrutinize” the details of an important aspect of the scene, especially in the cases where information delivered from the preattentive processes is not enough. Kowler and Anton (1987) conducted an experiment to show that if an object is not identifiable at a first glance, probably due to unfamiliarity, narrowed attention with fixations is required – in their experiment, alterations to the customary visual appearance of words, produced by changing letter order or orientation, slowed reading as saccades were made to look at every letter in sequence. Similarly, it has been reported that attending to the location of a change can overcome change blindness – i.e. changes were detected almost perfectly (Tse, Sheinberg, & Logothetis, 2003), and directing attention enhances visual perception of the corresponding location – e.g. lowered sensory thresholds, enhanced resolution, etc. (Bisley & Goldberg, 2003; Carrasco & Yeshurun, 2009; Kastner, Pinsk, De Weerd, Desimone, & Ungerleider, 1999).

Thus, our cognitive system utilizes both types of processes during perceiving a scene as those processes form an iterative cycle. It suggests that the perception of a visual scene is not only performed by focused attention but also constantly supported by nonattentive processes which provide the layout of an entity in a scene to which attention is directed. Since we proposed that visual perception happens at various levels of whole-part hierarchy, the currently attended entity might be an object, a part of an object, or even a group of objects in a certain relationship (e.g. the actor and the object in an action event) – recall the earlier discussion that the area of attentional focus need not be confined to a visual "object" but it can be defined in a much wider sense. The implication is that the layout provided from nonattentive processes would vary just as much, from a single object to the entire scene, depending on the coverage of current attention. In other words, nonattentive processes may provide the layout of not only the entire scene but also an event, a person, or an object that the person is holding, etc.

Therefore, even though the term "gist" has been typically used for addressing a holistic (covering the entire visual field) and fast (happening in about 100ms) recognition of a "scene" (Greene & Oliva, 2009; Oliva & Torralba, 2001), we need to define a broader sense of gist. The idea is that not only can we analyze the entire scene for gist, but also we can segment a part of the scene and get the gist of that – *gist works at all levels*. Perception of an entity, whether it a single part, an object, an event, or the entire scene, instantly produces a gist when recognized (by preattentive processes), and the successive examination (through focused attention) is guided by the gist which provides the layout of the entity (from scene/object schema, world knowledge, etc.). Note that sometimes instant recognition of an entity might not be possible, especially when the complexity of the entity is high or the entity itself is ambiguous. In that case, the gist (and the resulting layout) is not available until more thorough inspection of the entity, possibly with multiple attentional focuses, is carried out – and then it might not be appropriate to call it a "gist" anymore.

Conversely, the universality of gist also blurs out the distinction between a scene and an object, highlighting a unified scene perception process happening at various hierarchical levels. Thus, what is a scene at one level of analysis may be an object at another.

Although we propose a unified view of the gist for objects and scenes, their processing is not necessarily supported by the same neural circuitries. Rather, our proposal emphasizes the cognitive aspect of processing and the resulting conceptual framework benefited by such unified processes. In fact, a body of studies suggested a neural dissociation in processing objects and scenes. Encoding a layout of wide space has long been associated with the parahippocampal place area (PPA) – in fact, this is how it acquired such a name – and a number of studies reported that activity within this area selectively responded to spatial layout of a room or a scene (Epstein, et al., 2003; Epstein, et al., 1999; Epstein & Kanwisher, 1998). On the contrary, the lateral-occipital cortex (LOC) has been associated with recognition of objects and storing object templates in number of studies (Epstein, et al., 2003; Grill-Spector, Kourtzi, & Kanwisher, 2001; Peelen, Fei Fei, & Kastner, 2009). However, a recent study suggested that the gist of a scene may be processed by a similar early mechanism for the gist of an object (J. G. Kim & Biederman, 2010) by showing that object pairs shown as interacting (e.g. a bird perching on a birdhouse), compared with their side-by-side depiction (e.g., a bird simply put besides a birdhouse), elicited greater activity in the LOC.

2.7. Perception Beyond Fixations

In our proposal of hierarchical scene perception, in which attentive and nonattentive processes form an iterative cycle

through the levels of the whole-part hierarchy (Section 2.6), we highlighted the variety of scales that the attended entity can take – it might be a part of an object, a group of objects, an event, or even the entire scene. Each of these scales is characterized by the size of coverage and the degree of integrity for the required perceptual processes – e.g. narrow coverage with high integrity for perceiving a part of an object (fine detail), and wide coverage with low integrity for perceiving an event happening among people (coarse layout). This is a good deal of reminiscent of the zoom lens model (Eriksen & Yeh, 1985), in which attentional resources can be distributed over the visual field, but with low resolving power, or continuously constricted to small portions of the visual field with a concomitant increase in processing power. Here we propose an *attention window* to encompass the area occupied by the entity under the current attentional focus. An attention window delineates a certain area on a scene where (attentive and nonattentive) perceptual processes are performed to build a coherent representation of the entity in the area. Thus, the attention window acts as the unit of scene perception process, whose "receptive field" constantly adjusts according to the entity to which attention is directed. In fact, the entity perceived through an attention window comes with a variety of scales as attentional focus traverses up (zoom-out) or down (zoom-in) the perceptual hierarchy, and the depth (i.e. level of detail) and the size of an attention window vary accordingly – the attention window gets narrower and deeper as attention zooms in, whereas it gets wider and shallower as attention zooms out.

As other approaches consistent with the zoom lens model asserted (e.g. Castiello & Umilta, 1990), the key assumption of the attention window is that the area covered by visual attention is flexible. In fact, a number of studies provided evidence supporting flexibility in the size of visual attention. Müller and colleagues (2003), for example, provided neurophysiological data indicating the zoom-lens-like modulation on activity in multiple retinotopic visual areas (V1, V2, VP, and V4) in association with the size of the attended region. Rolls and colleagues (2003) demonstrated that the receptive fields of neurons in the inferior temporal cortex of the monkey brain differ according to the scene complexity. Hopf and colleagues (2006) showed that the receptive fields in visual areas (the LOC and V4) generally match the size of attending objects and they are adjusted rapidly (within 250~300ms) in response to moment-by-moment changes of scale. Moreover, results from behavioral experiments also suggested that subjects adjust the size of their attentional focus: depending on the task (Cave & Kosslyn, 1989; Oliva & Schyns, 1997), or according to the perspective scale (Jefferies, Gmeindl, & Yantis, 2011).

However, as Wright and Ward (2008) claimed that mental focus ("covert" attention) should be differentiated from the act of directing sense organs towards a stimulus source ("overt" attention), one should not equate an attention window with a fixation of eye gaze. According to Wright and Ward, (covert) attention is thought to be a neural process that enhances the signal from a particular part of the sensory panorama. An attention window defines an area of such neural enhancement of sensory perception (i.e. a receptive field), and it may comprise one or more gaze fixations. Many studies have suggested such a separation, especially the dissociation between neural systems for endogenous attention and oculomotor planning (Belopolsky & Theeuwes, 2009; Posner, 1980). Other studies also reported subjects' unawareness of involuntary saccades to unattended locations (Deubel, Irwin, & Schneider, 1999; Mokler & Fischer, 1999).

Figure 2.7-1: Examples of subscenes that emphasize various aspects of a scene. (A) covers the entire scene and it captures the layout of the filming event while (B) focuses on the hitting action happening between the woman and man, conveying more detailed information in a relatively smaller area. (C) is confined within the area of the woman while it delivers detailed properties of the area, such as woman's prettiness or the color of the dress.

The implication of the account on the attention window and the hierarchical vision process as discussed so far is that a visual scene is perceived neither as a whole nor as pieces, but rather as entities of various scales that are apprehended by attention windows of diverse size, shape, and depth. Given that the representation built from a scene can never be complete or in exhaustive detail, only the objects or events with cognitive significance are perceived through attention windows. In other words, the vision system selectively deploys attention windows to capture a certain aspect of the scene, which serves the current interests and goals. For example, for the question "*what is it John holding in his hand?*", a wider attention window is distributed over the scene to position John and his hand, and then a narrower attention window is employed to extract details of the object held in the hand. The set of entities perceived through such attention windows may form a scene where John is holding a book, which serves the current task of identifying the object held in John's hand. The formed scene represents the cognitively important (i.e. task-relevant) aspect of the scene at this moment, which is generally described in terms of an event structure, such as the event that *a book is held in John's hand*. We use the term *subscene* to describe such an aspect of the scene.

Itti and Arbib (2006) proposed a notion of "minimal subscene", which contains the "minimal" amount of information to describe a single action-related event – e.g. an amount of information to relate an agent, action, and a patient. Once an object or action has captured attention, it will act as an "anchor" to search for other related elements in the scene to complete a minimal subscene. One or more elements in a minimal subscene may in turn become an anchor for linking other elements in the scene, extending the minimal subscene into a bigger structure – an "anchored subscene". Itti and Arbib proposed a minimal subscene to be the basic unit of action event recognition, in which an agent interacts with objects or other agents.

However, the type of events we are dealing with extends beyond a simple action event since in the current framework, the SemRep has been proposed to be a middle-ground representation between the observation and description of a natural scene. This necessitates the extension of the notion of minimal subscene. Therefore, we define a subscene as a more general construct that captures a cognitively significant event of various types during scene perception. A subscene is a partial view

of the scene covered by entities – agents and/or objects that are linked via actions and/or other relationships – that are delimited (and perceived) by one or more attention windows. Within the framework of action-oriented perception, we propose that the vision system perceives a scene in terms of subscenes that come in various event types and covering areas, representing a particular view on the scene at a certain moment.

Depending on the size and depth of attention windows involved in forming a subscene, qualitatively different descriptions of the scene are possible, from "*a fist bumps on a face*", to "*the woman is hitting the man*", or even to "*people are fighting*" (B in Figure 2.7-1). This means that information contained in two subscenes may vary even if they cover the same area. Holsanova (2008) also emphasized such a subtlety in scene perception as she claimed that concrete objects can be viewed differently (at different levels of specificity) on different occasions as a result of our mental zooming in and out.

By the nature of subscenes which take diverse event structures and coverage areas that extend from a part of an item to the entire scene, a conceptual hierarchy may be formed among subscenes – i.e. a subscene may be conceptually "subordinate" to another subscene. When the vision system builds a SemRep of the perceived scene, the event structures captured within subscenes are hierarchically organized according to their cognitive importance, such as task relevancy, perceptual salience, or temporal arrangement. Such an organization results in a SemRep, which can too be viewed as representing a subscene as a whole, being divided into a number of substructures, each of which represents a particular event of the scene delivered through a subscene (see Figure 2.7-2 for an example). The hierarchical structure of a SemRep may influence the description process by imposing priority to the event ranked at a higher level (see Section 4.4 for detailed processes of utterance production).

Moreover, given that a subscene represents a cognitive construct of a coherent event structure, the hierarchical organization of subscenes may subserve the chunking mechanism of visual memory. A subscene may be treated as a flexible unit of cognitive processing, especially the process of scene perception and description, which may exploit the mechanism of chunking in accessing and storing perceived visual items. In fact, Cowan (2000) claimed that during short-term memory tasks, a coherent scene (or a chunk) is formed in the focus of attention and that scene can have about four separate parts (also as chunks) in awareness at any one moment while the focus of attention shifted back and forth between the hierarchical levels of these scenes. But further research is required for a clearer view on this account.

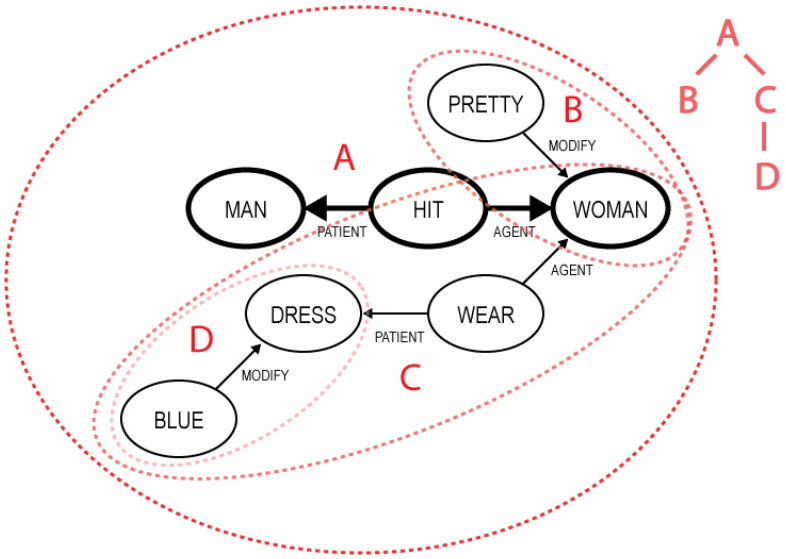


Figure 2.7-2: A SemRep possibly generated by the scene shown in Figure 2.3-3. A particular event of *a woman hitting a man* is represented in the whole subscene, which involves four sub-level event structures that are also represented as subscenes. These structures are hierarchically organized according to their cognitive significance with (A) at the top rank (depicted with thick lines). (A) is treated as the "main event" of the whole subscene, resulting in such produced sentences as "*a pretty woman in blue is hitting a man*", "*a pretty woman who's wearing a blue dress is hitting a man*", or "*a man is hit by a pretty woman wearing a blue dress*", etc.

There are basically three types of attentional procedures: *zooming-in*, *zooming-out* and *shifting*. When zooming-in, the attentional focus traverses down the perceptual hierarchy, making the attention window narrower and deeper (extracting fine details), whereas when zooming-out, the attentional focus traverses up the hierarchy, making the attention window wider and shallower (scanning a coarse layout). Shifting does not involve traversing the attention hierarchy but moves attention window to another location. The vision system deploys these procedures to build a SemRep in a manner depending on the scene property and task goals.

For all the possible steps that a subscene is perceived and encapsulated into a SemRep, we propose two broad scenarios as illustrated in Figure 2.7-3. The first scenario (perception by specification) covers the types of perception process where the gist is specified first and the successive fixations add more details within the boundary of a scene representation, whereas the second scenario (perception by extension) covers the types of perception process where a scene representation is built by extending its boundary as more information is perceived. In real cases, scene perception process may combine the two.

We now examine in detail how a subscene is perceived in either of the basic cases, as illustrated in Figure 2.7-3. We assume that an attention window is initially placed on the most perceptually and cognitively salient area, which is in this case, about the area of the hitting event between the man and the woman.

In the first case (A in Figure 2.7-3), the event perceived through the initial attention window is assumed to be clear enough to provide a gist, or a layout, of the event, which is a transitive action event, although specific details are not yet provided (recall *gist works at all levels* in Section 2.6). Note that the event structure (with missing components) has been already specified at the first stage. Thus, the vision system can be guided by the layout to fill in missing components in the

event structure. At the second stage, an attention window has been placed on the woman's face (zooming-in) to identify the actor of the action event. In this case, various factors might come into play in selecting the component to attend to, such as perceptual salience (e.g. the woman's face is most prominent among other components in the subscene), linguistic bias (e.g. an active sentence requires the actor to be produced first; Section 4.5), or cognitive preference (e.g. the actor is generally the most significant component in an action event). At the third and the fourth stage, the missing components, the hitting action and the man, have been identified consecutively (shifting), completing the subscene of the hitting event between the man and the woman. In this case, the subscene is perceived by "specification" since the layout is initially available and the successive stages specify the details of the subscene.

On the other hand, in the second case (B in Figure 2.7-3), the event perceived through the initial attention window is assumed to be *not* clear enough to provide a layout of the event. Thus, at the first stage, nothing has been specified out, so the vision system focuses on more specific entities in the scene. At the second stage, an attention window has been placed on the man's face (zooming-in), resulting in creating a node for the man. Currently, the subscene only contains the man node as its existence in the scene is the main event of the subscene. The perception of the man's face leads to the perception of the fist, thus resulting in placing an attention window over the fist (shifting) at the third stage. At this stage, the hitting action has been identified and the man is specified as the patient of that action, extending the subscene to now contain the passive action event of the man being hit. At the fourth stage, the actor of the action has been identified as an attention window has been placed on the woman's face (shifting), completing the subscene of the hitting event between the man and the woman. In this case, the subscene is perceived by "extension" since the subscene is built incrementally as more components are perceived.

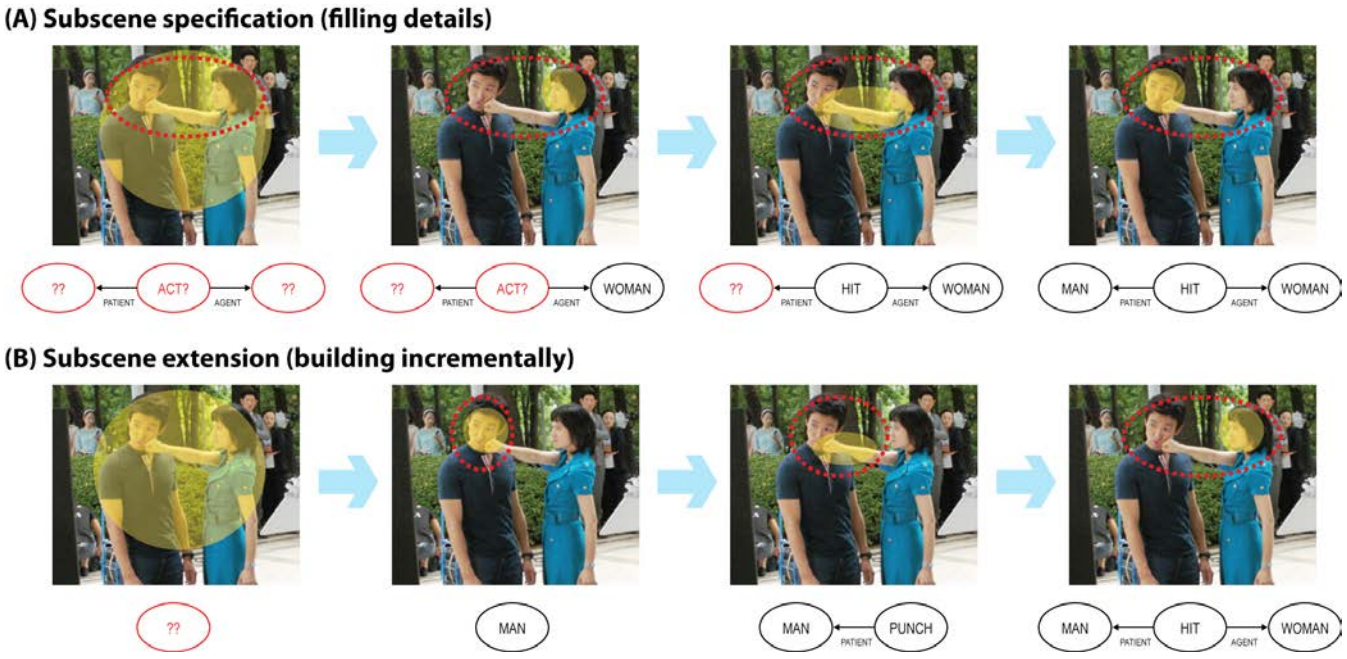


Figure 2.7-3: An illustration of two cases of (sub)scene perception. (A) illustrates the case of "specification" where a subscene is perceived by filling in details, whereas (B) illustrates the case of "extension" where a subscene is perceived in an incremental manner. Yellow ovals represent attention windows, red ovals represent subscenes, and red nodes represent detected but unidentified entities of the subscene. See text for more detailed description.

The key difference in the earlier examples is the availability of an immediate layout. Especially during complex and natural scene viewing, only a limited amount of information is carried across saccades (e.g. Henderson & Hollingworth, 1999; Hollingworth & Henderson, 1999), leading to temporally dissociative (but not so clearly) processes for perceiving an abstract layout and inspecting missing object details. The layout availability and the subsequent temporal dissociation in perception have a significant implication in the process of scene perception and description as described in the detailed exposition in Chapter 5. The interplay between the layout availability and the resource constraints given to the system will drive the system to produce various patterns of gaze fixation and produced utterance. However, as far as we know, there is unfortunately no direct data addressing this issue of the layout availability under different conditions – see Section 5.2 and Section 5.3 for relevant experimental findings.

2.8. Event and Episode Structure

Although it is not rigorously considered in the current work, it should be worth addressing some of the implications of episode structure in the framework of SemRep since the notion of subscene (Section 2.7) has an intrinsic relationship with an episode structure. A subscene delineates a cognitively significant event, which can be interpreted as a type of episode, especially when a number of subscenes are formed over time. More specifically, a series of snapshots of subscenes being formed over a certain temporal period may be concatenated into a series of event structures, each of which may present an episode. Recall that subscenes are hierarchically organized and one of the criteria for the hierarchy is temporal order. In the case of episodes, a causal relationship might also be considered as such a criterion.

Moreover, the SemRep has been originally proposed to be used not only for static scenes but also dynamic scenes. This eventually requires a SemRep to represent a series of events happening over a certain time period while they actually form a type of a story. Rumelhart (1977) argued that a story is encoded as an episodic structure or an event that could be traced to an "actor-action-goal" sequence, and this is quite consistent with the underlying idea in the proposal of the notion of minimal subscene – a minimal subscene defines an action event representation involving an agent interacting with objects or other agents. Although it goes beyond the range of the current work, the SemRep should eventually factor into such an episode structure encapsulated for an action or more general event. A group of those episodic structures could be connected by links expressing various types of spatial, temporal and causal relations.

Although there have been efforts among philosophers, linguists and psychologists to develop a classification of event types that accurately captures logical entailments (e.g. Zacks & Tversky, 2001), they still appear far from converging on what constitutes an event. However, perceiving boundaries of events seems to be somewhat universal. It has been suggested that there is a significance overlap in detecting action and movement segments during perceiving action sequences (Hanson, Hanson, Halchenko, Matsuka, & Zaimi, 2007), and the concept of event seems to be already developing in a very early age as 10-month-olds could distinguish relevant elements in events involving "giving" and "hugging" (Gordon, 2003). Moreover, it has been suggested that spatiotemporal, rather than specific sensory (e.g. shape), information is necessary in forming episodic object representations (Henderson, 1994).

Burgess and colleagues (2001) and Shastri (2002) asserted the role of the hippocampus system in storing episodic

memory by highlighting its computational role as a “content-addressable” associate memory. According to their claims, the hippocampal system is the storage unit for patterns, which can retrieve complete patterns from partial cues. In their models of episodic memory, memory traces persist in the hippocampal system as long as remembered, and these memory traces are represented and accessed by implicit neural circuits that are rapidly formed inside the hippocampus, as a result of long-term potentiation (Shastri, 2002) or Hebbian learning (Burgess, et al., 2001). The former focused on the relational event structure whereas the latter emphasized the spatial context or configuration as the key component of the memory. Nyberg and colleagues (1996) also suggested the existence of general encoding and retrieval networks of episodic memory centered around the hippocampus.

Moreover, areas around the perirhinal cortex were also suggested to be associated with longer mnemonic performance in visual recognition tasks. Patients with complete damage to the perirhinal cortex exhibited intact visual recognition capability for immediate memory span (0 ~ 2 sec.) but showed impaired performance for a longer period of time (more than 25 sec.), which was much worse than other amnesic patients with lesions in different brain areas (Buffalo, Reber, & Squire, 1998).

Although both remembering events (episodes) and remembering object semantics require memory for a longer duration, there appears a qualitative distinction between episode recall and semantic recognition. Aggleton and Brown (1999) argued that impairment in encoding and recalling of episodic memory (anterograde amnesia) is due to damage in the hippocampal system whereas familiarity judgment reflects an independent process that depends distinctly on the perirhinal cortex. Similarly, Vargha-Khadem and colleagues (2001) also argued such a distinction between recollection-based versus familiarity-based judgments, while reporting that patients with hippocampal pathology showed severe impairment in episodic memory (recall) although their semantic memory (recognition) was relatively preserved.

Thus, the distinction between storing episodic information and object semantics suggests that in order to fully handle episode structures, the neural substrates for the SemRep eventually need to extend beyond the network of concept and semantics (Section 3.1) and the network of visual perception (Section 3.2) to include the network of the hippocampal cortex and the adjacent areas. These areas were claimed to encode abstract multi-modal information (Epstein, et al., 2003; Epstein, et al., 1999; Epstein & Kanwisher, 1998; King, et al., 2002; O'Keefe, 1999; Pierrot-Deseilligny, et al., 2002), which is appropriate for representing event structures of episodic memory as they intrinsically require multi-modal and abstract encodings of object representations.

Chapter 3. Integrative Framework of Vision and Language

3.1. Network of Semantics and Concepts

While the layout of the perceived scene is captured by the graphical components (i.e. nodes and edges) of SemRep, the more specific semantics of the entities or their relationships are represented by the "concepts" that are tagged with the corresponding nodes and relations. Although the current work implements the computation of concepts through simulation of a symbolically represented schema network rather than through simulation of the brain's neural networks, we propose that each concept is associated with one or more perceptual schemas whose processing is claimed to be instantiated in neural activities.

A number of lesion studies on conceptual knowledge reported that the loss of conceptual knowledge of various categories follows a certain topographical pattern, implying that the knowledge of various concepts is distributed all over the brain in a category-specific manner. Especially, the conceptual knowledge of objects was reported to be generally structured around two primary categories of living things (e.g. animals or plants) and non-living things (e.g. tools or artifacts) (Caramazza & Shelton, 1998; Lambon Ralph, Lowe, & Rogers, 2007; Tyler & Moss, 2001; Warrington & Shallice, 1984). Gainotti (2000) sought the reason of such topographical distribution of conceptual knowledge from distinctive brain areas associated with the process of corresponding categorical properties, arguing that the category-specific disorder is crucially related to the kind of semantic information processed by the damaged areas – deficits in living things are due to lesions in sensory and perception areas (e.g. the inferior temporal cortex) as the concepts of animals are more dependent on their perceptual features (mostly visual) whereas deficits in non-living things are attributed to lesions in the motor-related areas (e.g. the fronto-parietal cortex) as the concepts of non-living things are defined more in terms of their functional properties. This categorical distinction in conceptual knowledge, especially between animals and tools, has been further supported by the studies based on linguistic tasks (e.g. naming) on concrete objects (Beauchamp & Martin, 2007; Chouinard & Goodale, 2010; H. Damasio, et al., 1996; H. Damasio, et al., 2004; A. Martin, Wiggs, Ungerleider, & Haxby, 1996).

Moreover, a number of studies also reported dissociation between linguistic processes of object words (typically nouns) and action words (typically verbs), arguing for neural separability between object concepts and action concepts (A. R. Damasio & Tranel, 1993; Pulvermüller, Mohr, & Schleichert, 1999; Vigliocco, Vinson, Druks, Barber, & Cappa, 2010). In another study, the lesions of subjects with impaired retrieval of conceptual knowledge for actions showed the highest overlap in the distinct brain regions for action-related processes, such as the left premotor/prefrontal, the left parietal, and the posterior middle temporal regions (Tranel, Kemmerer, Adolphs, Damasio, & Damasio, 2003), which are different from the brain regions that are generally associated with deficits in the concept of concrete objects. Furthermore, it has been reported that linguistic tasks on verbs or action-related sentences correlate in activation of the motor and premotor cortex (Kemmerer, Gonzalez-Castillo, Talavage, Petterson, & Wiley, 2008; Tettamanti et al., 2005) while the activation was suggested to happen in a somatotopic fashion (Buccino et al., 2001; Hauk, Johnsrude, & Pulvermüller, 2004) – for example, reading action words referring to face, arm, or leg actions (e.g., "to lick", "pick", or "kick") or observation of such actions differentially activated areas along the motor strip that either were directly adjacent to or overlapped with areas activated by performing actual

movement of the tongue, fingers, or feet. This line of studies suggest that the neural distinction in conceptual knowledge is not limited to the categories of concrete objects but also extended to more general categorical levels, such as action and object. Also, they further support the claim that conceptual knowledge is topographically distributed and grounded in the actual neural circuits that process the corresponding concepts in the brain.

However, one should note that conceptual knowledge is not necessarily modality-specific although it is grounded in the neural areas of specific sensory and motor processes. Rather, it was suggested that semantic memory consists of both modality-specific and supramodal representations, the latter supported by the gradual convergence of information throughout higher-order association areas (Binder & Desai, 2011). Such areas are a type of *convergence zones* (A. R. Damasio, 1989), which are proposed as neurally manifested spaces where stimulus patterns of feature maps in primary and early sensory/motor cortices are integrated and shared across different modalities, and such shared features are captured by *conjunctive neurons* in convergence zones. Convergence zones are claimed to be hierarchically organized as the conjoined features of instances of a certain category level are captured by conjunctive neurons in the convergence zone of that level, whose commonalities are successively captured by conjunctive neurons of another higher-level convergence zone, thus establishing a foundation for abstract concept representation (Simmons & Barsalou, 2003).

Large regions of frontal, temporal, and inferior parietal cortex are suggested to be such higher-order association areas. Especially, a body of studies suggested that the anterior temporal pole is a place where increasingly abstract representations are stored (H. Damasio, et al., 1996; H. Damasio, et al., 2004; A. Martin & Chao, 2001; Murray & Richmond, 2001; Rogers, et al., 2004; Vargha-Khadem, et al., 2001). Emphasis has been given on the temporal pole based on the neuroanatomical evidence that massive multimodal inputs converge on the anterior medial temporal regions, particularly the perirhinal cortex, forming a caudal-rostral gradient within the temporal lobes (A. Martin & Chao, 2001). Moreover, Rogers and colleagues (Rogers, et al., 2004) proposed that the regions act as a cross-modal “hub” where modality-specific perceptual, linguistic and motor representations communicate with one another. Other studies highlighted the inferior parietal cortex and relatively wide areas of the inferior and middle temporal cortex as possible candidate sites for higher-level convergence zones (Binder & Desai, 2011; Chouinard & Goodale, 2010; Desai, Binder, Conant, Mano, & Seidenberg, 2011).

As our work should be interpreted within a bigger framework of the coupling of scene perception and speech production, we highlight the mechanisms for integration of perceptual and motor schemas across systems of various modalities and abstraction levels. The implication of the neural substrates supporting concepts and semantics as outlined so far is that the SemRep, more specifically the “concepts” of the SemRep, are based on well-localized representations of objects and actions while higher-level association areas (or convergence zones) provide bidirectional links between those representations, allowing formation of abstract conceptual representations with cross-modal properties. For instance, the perception of an apple may involve invoking a number of perceptual schemas of various modalities, such as visual, tactile or olfactory systems. These activated perceptual schemas are then integrated over associative areas to form a relatively abstract schema of an “apple”, which may be in turn encoded within the concept of the SemRep – recall that a SemRep is basically proposed as an abstraction of schemas for perception of a particular aspect of the current scene.

If the current plan of action requires more information, conversely, further activation of schemas may happen through the cross linkages of such associative network. Perception of an object for a particular course of action might invoke schemas

that are not available from the immediate perception – e.g. perceiving of an apple for eating may invoke motor schemas for jaw movements or perceptual schemas of how an apple tastes or so. Conceptual representations of an abstract level, like the ones encoded within the SemRep, can be "fleshed out" by invoking lower-level schemas of modality-specific representations when more detailed information is required. For example, Desai and colleagues (2011) reported a case where the involvement of sensory-motor systems in metaphor understanding showed an inverse-proportional relation with the familiarity of the metaphors – detailed simulations are used for understanding unfamiliar metaphors while these simulations become less detailed as familiarity and contextual support increase. They proposed the anterior inferior parietal lobule as a high-order interface between sensory-motor and conceptual systems.

In both cases where more concrete representations are integrated to yield more abstract representations and processing of such abstract representations are supported by those concrete representations, higher-order association areas act as a neural foundation for integrating schemas across systems of various modalities and abstraction levels, laying a ground for categorical concepts.

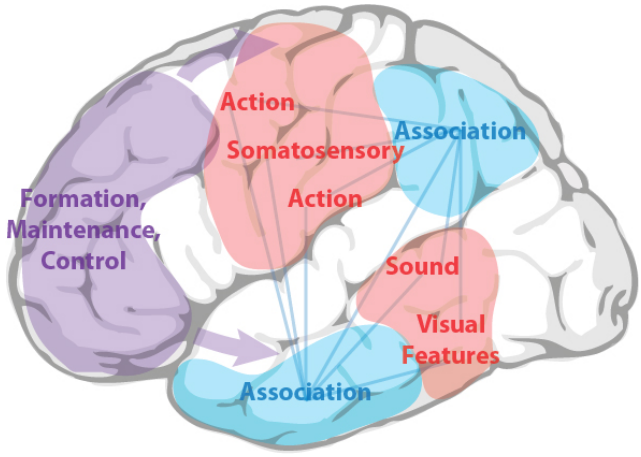


Figure 3.1-1: Modality-specific sensory and action systems (red) provide detailed representations shaped through experiences while high-level temporal and inferior parietal association areas (blue) store increasingly abstract conceptual representations. These two types of areas form a LTM network provides a means to categorical access of conceptual entities. Frontal regions (purple) control the goal-directed activation and selection as well as maintenance of the information stored in those temporo-parietal cortices, forming a WM network of concepts and semantics.

The mechanisms of schema activation mediated through association areas, as outlined so far, have implications of a particular working memory (WM) system for conceptual representations. In his account of "active memory", Fuster (1997) emphasized the transient activation of the associative network of perceptual memory fragments that are reactivated from networks of long-term memory (LTM). Similarly, Ericsson and Kintsch (1995) proposed a type of working memory system which is extended to include storage in LTM while they viewed information kept in WM as a type of "retrieval cue" through which the knowledge acquired from experiences or activities can be directly accessible from long-term memory. Inspired by this, moreover, a specific form of working memory system has been proposed that the contents of WM are understood as "activated" representations from LTM, which are currently within the focus of attention (Cowan, 1999; Oberauer, 2002).

Since all of these studies share the view point that WM is established by an activation of patterns stored within LTM, they are intrinsically consistent with the stance we take in the framework of schema theory, where a dynamic assemblage of schema instances that are temporarily activated from LTM forms the WM of an organism. The WM contains a set of schemas whose activation and de-activation continuously happen according to the dynamically changing action goals and needs of the organism. Different types of schemas in various degrees of abstractness are accessed by the spreading of activation in the network of semantics and conceptual knowledge, which is neurally grounded in sensory and motor systems of different modalities distributed over the brain, while higher-order association areas act as the network hub.

However, we would also like to emphasize that WM does not consist of simply activated schemas, but rather it consists of "activated and parameterized" instances of schemas. As Baddeley (2003) emphasized during his insist on the necessity of the episodic buffer, simply activating representations within LTM seems insufficient, especially when manipulating and creating new representations are required. The representations stored in LTM provide prototypes of newly established representations whose specifics are defined by parameters tuned to particular situations. This requires WM to maintain active copies of schema instances throughout the processing, a type of dynamic representations of concepts and semantics, such as the SemRep.

We propose that this semantic WM system is established around the frontal network that stretches to temporo-parietal regions. While wide regions in temporal and parietal cortices play a role as LTM for storing various abstraction levels of conceptual representations, frontal areas are reported to be responsible for "top-down" mechanisms to mediate retrieval of such representations and their maintenance (Ishai, Ungerleider, Martin, & Haxby, 2000). Especially regions around the pars orbitalis (BA 47) and the pars triangularis (BA 45) in the ventrolateral prefrontal cortex (VLPFC) are claimed to be involved in semantic retrieval, recollection of contextual details, and resolution of interference in WM and task switching (Badre & Wagner, 2007). More specifically, it has been reported that BA 47 is responsible for facilitation of semantic information and BA 45 is involved in sematic selection processes (Gold et al., 2006).

On the other hand, the dorsolateral prefrontal cortex (DLPFC) has been claimed to contribute to manipulating associations among semantic items activated in WM by strengthening (Blumenfeld & Ranganath, 2006) or weakening (Chee, Sriram, Soon, & Lee, 2000). These data suggest that anatomically separable subregions within the lateral PFC may be functionally distinct and are consistent with models that posit a hierarchical relationship between dorsolateral and ventrolateral regions such that the former monitors and selects goal-relevant representations being maintained by the latter (Wagner, Maril, Bjork, & Schacter, 2001). Regions in the posterior parietal cortex (PPC), furthermore, may support an active WM buffer by dynamically directing attention to internal and mnemonic representations that are dependent on the medial temporal lobe (Wagner, Shannon, Kahn, & Buckner, 2005).

The type of WM system for conceptual representations proposed in the current work is illustrated in Figure 3.1-1. The LTM network of semantics and concepts consists of neural areas of specific sensory and motor processes as well as association areas for cross-modal integration and abstraction. Schemas of various modalities and abstraction levels are activated (and parameterized) from the LTM network while top-down bias signals from the frontal and parietal areas mediate and maintain the assemblages of those schemas, yielding a SemRep.

3.2. Network of Scene Perception

Although it is probably an over-simplification of the true state of affairs in the visual cortex, the perception task of a visual scene was claimed to be basically comprised of "locating" and "identifying" (Ingle, Schneider, Trevarthen, & Held, 1967). The dichotomy of locating and identifying was later linked to primate cortical anatomy in the work of Mishkin & Ungerleider (1982) who distinguished two streams of visual process in the striate and extrastriate cortex in the monkey brain, and since then, the idea of separate pathways for visual processing has been widely accepted.

In primates, both streams originate from the primary visual cortex (V1/V2) but one extends ventrally from V1 through V4 to the inferior temporal (IT) cortex while the other extends dorsally from V1 to the posterior parietal (PP) cortex. The former stream is generally assumed to subserve object recognition (i.e. identifying) whereas the latter is characterized as mediating spatial memory (i.e. locating). Therefore, the ventral pathway is called the "what" pathway since lesion to this pathway in monkey impaired the performance of visual pattern discrimination and recognition but not object location tasks. On the other hand, the dorsal pathway is called the "where" pathway since quite the opposite results were observed in monkeys with lesions to this pathway (Mishkin, Ungerleider, & Macko, 1983). Moreover, it has been suggested that similar to nonhuman primates, multiple visual areas in the cortex of the human brain are organized into two functionally specialized and anatomically segregated processing pathways (Ungerleider & Haxby, 1994).

Most of the evidence supporting this dichotomy comes from visuo-spatial WM studies. Wilson and colleagues (1993) segregated WM components for the spatial location of visually presented objects and the visual characteristics of those objects – WM for the spatial location involves the posterior parietal cortex (PPC; where spatial vision is processed) and its connections with the dorsolateral prefrontal cortex (DLPFC), while WM for object characteristics relies on connections between the inferior temporal (IT) lobe (where object features are processed) and the inferior convexity of the prefrontal cortex (PFC). Similarly, Sala and his colleagues (2003) directly compared the patterns of response during WM tasks for face identity, house identity, and spatial location, and reported that the superior PFC produced the greatest response during spatial WM tasks while the middle and inferior PFC produced the greatest response during object WM tasks, independent of the object type. Finally, Irwin (2004) reported that cognitive operations requiring visuo-spatial processing (e.g. mental rotation) were suppressed during saccades while saccades did not interfere with stimulus recognition and identification tasks. All of these studies suggest the dorsal-ventral functional segregation for spatial and non-spatial information.

The dorsal stream has been reported to be grounded in the regions for guiding saccadic movements and visual attention deployment. More specifically, the dorsolateral part of the PFC, including the frontal eye fields (FEF), and a portion of the PPC were suggested to be significantly involved in guiding eye movements and attention deployment (Curtis, 2006; Curtis & D'Esposito, 2003; Dominey & Arbib, 1992), implying a tight relationship of these regions with visuo-spatial processes. In fact, a number of studies emphasized the primary role of these frontoparietal regions in spatial memory (Ma, et al., 2011; Ma, et al., 2004; Ma, et al., 2003; Pierrot-Deseilligny, et al., 2002; Sawaguchi & Iba, 2001).

On the other hand, the ventral stream has been reported to be based on the neural circuitry for encoding and retrieving visual features and identity of objects. Especially, strong connectivity between the PFC and the IT cortex has been implicated in object memory processes, and it has been suggested that the PFC and the IT cortex forms a WM circuit where the PFC is a source of feedback inputs to the IT cortex, biasing activity in favor of behaviorally relevant stimuli (Desimone, 1998; Miller

& Desimone, 1994; Miller, Erickson, & Desimone, 1996; Ranganath, DeGutis, & D'Esposito, 2004).

Two distinct types of information conveyed through these separate streams presumably join at frontal areas. During WM tasks, neurons in PFC showed both object-tuned ("what") and location-tuned ("where") delay activity (Rainer, Asaad, & Miller, 1998; Rao, Rainer, & Miller, 1997; White & Wise, 1999), suggesting that the perceived visual information – spatial and non-spatial – is integrated in the PFC. Moreover, a number of studies suggested that there is dorsal-ventral segregation in processing visuo-spatial information even within this area – the dorsolateral part of the PFC is responsible for processes in spatial information while the ventrolateral part is more involved in selection and retrieval of object identity and features (Munk et al., 2002; Ninokura, Mushiake, & Tanji, 2004; Sala & Courtney, 2007; Sala, et al., 2003). This spatial and non-spatial dissociation may be "multisensory" and may be applicable beyond the vision system to the systems of more general domain, such as the auditory system (Romanski, 2007). However, the implication of functional dissociation in the PFC is somewhat controversial (e.g. Rao, et al., 1997; White & Wise, 1999). In any ways, the PFC seems to be crucial in integrating visual information of different characteristics and forming a coherent scene representation.

The evidence on visuo-spatial WM reviewed so far has implications on a particular role of the PFC in sustaining visual representations. As mentioned in the account of the dorsal stream above, the frontal-parietal network is implicated in sustaining spatial memory. Curtis and his colleagues (2005), for example, claimed that the network of the FEF, the dorsolateral PFC, and the PPC supports spatial WM by sustaining covert attention at a particular location. Simiarly, Sawaguchi and Iba (2001) argued that specific visuo-spatial coordinates are represented in a topographical memory map in the DLPFC. Although regions in both the PPC and the PFC were associated with WM, evidence suggests that the PFC plays a prime role in preserving and maintaining processes of visuo-spatial representations as the PPC is more involved in providing a capacity-limited store (Qi et al., 2011; Todd & Marois, 2004). In fact, it has been reported that that only regions in the PFC (regions in the FEF) showed sustained delay period activity for both of the working memory and the attention task while the PPC (areas around the intraparietal sulcus) did not show any delay period activity (Offen, Gardner, Schluppeck, & Heeger, 2010).

As briefly mentioned in the earlier account on the ventral stream, moreover, the network formed between the PFC and the IT cortex has been suggested to be an object memory circuit where bias from the PFC works in favor of behaviorally relevant stimuli. Desimone and Duncan (1995) highlighted the similarity of top-down mechanisms in both object and spatial selection, and later Desimone (1998) reported that biasing of IT neurons in a WM task was remarkably similar to the biasing effects on the extrastriate cortex during visual search and spatially directed attention. Both of the studies pointed out the PFC as a main source of top-down feedback.

Moreover, the PFC also has been associated with selection processes in visual perception. Patients with unilateral frontal brain damage exhibited greater difficulty in shift from one aspect of an ambiguous figure to the other than did normal subjects (Ricci & Blundo, 1990). Similarly, activity in frontal and parietal regions was claimed to be associated with perceptual alternation during a phenomenon called binocular rivalry, which happens when dissimilar images are presented to the two eyes simultaneously (Lumer, Friston, & Rees, 1998), or conscious change perception (of human faces) (Turatto, Sandrini, & Miniussi, 2004). Moreover, a number of studies reported that the amount of cognitive load (as inflicted by requiring subjects to memorize a sequence of digits or to subtract numbers, etc.) showed a strong correlation effect with

perceptual selection processes (de Fockert, Rees, Frith, & Lavie, 2001; Pinsk, Doniger, & Kastner, 2004; Schwartz et al., 2005; Spinks, Zhang, Fox, Gao, & Tan, 2004; Yi, Woodman, Widders, Marois, & Chun, 2004); increasing cognitive load, which is followed by increased prefrontal activity, was associated with the behavioral performance of a selective attention task where subjects are required to ignore a distractor stimulus presented to the periphery of a visual field, thus implying that the PFC plays a significant role in perceptual selection (see Lavie, 2005 for more detailed review on the relationship between cognitive load and attentional selection).

Therefore, the role of the PFC seems crucial in visual perception processes – it is involved in perceptual selection, integrating visual information, and preserving perceived representations.

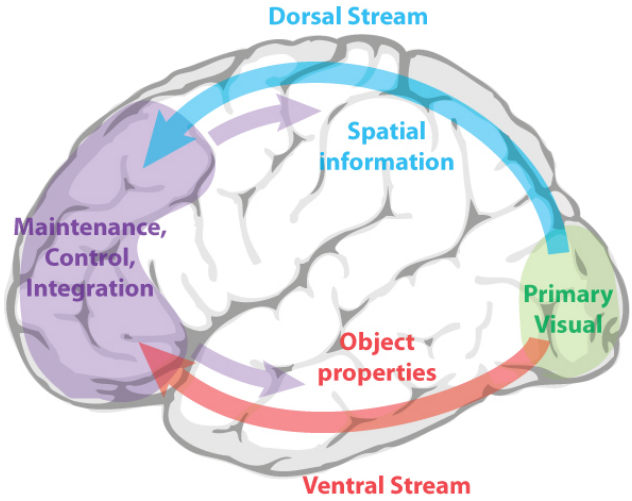


Figure 3.2-1: A schematized illustration of a WM network of visual scene perception with the two visual streams highlighted. The dorsal stream (blue arrow) is associated with spatial information, such as location, while the ventral stream (red arrow) is associated with non-spatial properties. Both of the streams start from the primary visual areas (green) and conjoin at PFC areas (purple). The PFC plays a primary role within this network by selecting and integrating different types of visual information and mediating a coherent scene representation.

Given all the implications of the PFC's primary role in visual perception, we propose a visuo-spatial WM network for visual scene perception which comprises areas in the dorsal and ventral pathway that are centered around the PFC (Figure 3.2-1). The spatial information (i.e. location) and non-spatial information (i.e. identity and properties) of visually perceived objects and entities are processed and delivered through the dorsal and ventral stream, then they are integrated within the PFC, forming a unified representation of a visual scene. In our framework of SemRep, the former is captured within the graph component (i.e. node and relation) while the latter is encoded within the associated concept.

This coherent representation of the perceived scene is "dynamically" built and maintained in the form of the SemRep within the network of visuo-spatial WM. Emphasis has been given to the dynamicity of the process since we view the SemRep as a very active representation which keeps changing over time, even for a static scene (as emphasized in Section 4.3). Scene perception is dynamic in its nature as only entities of enough cognitive significance (e.g. by task relevance, or perceptual saliency, etc.) would be perceived to build components (nodes or relations) of the SemRep while some

components may be changed or removed as more information is perceived or their activity level diminishes due to the loss of task relevance or the temporal decay.

Moreover, although higher-level cognitive areas and their vision- and memory-related processes are of our main focus, we would also account for their influences on earlier perception areas and their functions. Especially, evidence suggests that visual perception can be guided and even enhanced by the top-down bias from higher areas of memory and attention (Treue, 2003). For example, it has been reported that memorizing the shape of the target or target templates enhances performance in a visual search task (Oh & Kim, 2003; Soto, Humphreys, & Heinke, 2006) and directing attention enhances visual perception, such as resolution (Bisley & Goldberg, 2003; Carrasco & Yeshurun, 2009; Kastner, et al., 1999). This top-down bias is presumably inflicted in a hierarchical manner organized along the back-to-front axis in the vision system (Grill-Spector & Malach, 2004).

Therefore, the visuo-spatial WM network we are proposing here is not mere a storage buffer for visual representations but rather a complex of visual perception and dynamic maintenance mechanism. As Levedev and his colleagues (2004) argued, the PFC's function may go beyond the simple storage of visuo-spatial representations to include aspects of attention, such as the monitoring and selection of information. In fact, the overlap between the current WM network of scene perception with the earlier-proposed WM network of semantics and concepts (Section 3.1) does not appear coincidental. In both of the WM networks, the PFC (ventrolateral portion) and the IT cortex are involved in processing object-related representations, and the regions involved in attention control, the PFC (dorsolateral portion) and the PPC, support maintenance processes. This may suggest an integrative framework of WM for building and maintaining the SemRep where the PFC plays a prime function in control and mediating processes (see Section 3.4 for more detail on this account).

In addition to the areas in the scene perception network delineated so far, some studies suggested the rostral part of the superior temporal cortex (STC) as a site for multimodal sensory convergence for both object-related and space-related information, addressing its polysensory projections from both ventral and dorsal streams (Karnath, 2001; Thiebaut de Schotten et al., 2005).

3.3. Network of Linguistic Processes

Since the earlier studies on aphasic patients pioneered by Paul Broca and Carl Wernicke, it has long been postulated that perisylvian regions are deeply involved in linguistic processes, forming the "language network". Keller and colleagues (2001) suggested that the language process requires extensive collaborations of multiple areas in those regions – syntactic processing requires coordinated communication between Broca's and Wernicke's area, and phonological processing requires interactive communication among Broca's area, Wernicke's area, and the left inferior parietal lobule (IPL). In a recent review on fMRI studies in speech comprehension and production, Price (2010) summarized that activation was found in a wide area of bilateral temporal regions and the left frontal regions, including the left middle frontal cortex and motor and premotor cortex, for speech comprehension and production. Moreover, results from a sentence comprehension task (Stowe, Withaar, Wijers, Broere, & Paans, 2002) indicated that regions around the temporal and inferior frontal gyrus (on the left hemisphere) are responsible for sentence processing. Subcortical structures, such as the thalamus and striatum, were also suggested to be involved in language processing, especially in modulating the integrated representation of meaning at the sentence level

(Dominey, Inui, & Hoen, 2009).

An interesting aspect of the language network addressed in the current work is the establishment of the dichotomous view, similar to that of the vision network outlined earlier (Section 3.2). Landau and Jackendoff (1993) argued that a nonlinguistic disparity between the representations of "what" and "where" underlies how language represents objects and places. Wu and colleagues (2008) similarly suggested that languages consistently distinguish the path and the manner of a moving event in different constituents while such segregation respects the "dorsal-where and ventral-what" organizational principle of vision.

In fact, many scholars have made efforts to relate separate processes for syntax and semantics to such dichotomy. For example, Ullman (2001, 2004) claimed a dissociation of the neural substrates for lexical knowledge and syntactical processing, where the former is largely rooted in the temporal lobe whereas the latter depends on the procedural memory stabilized and automatized throughout cortico-striatal networks (frontal, basal-ganglia, parietal and cerebellar structures). This claim was further supported by a study on aphasics (Ullman, Pancheva, Love, Yee, & Swinney, 2005). Similarly, Piñango and Zurif (2001) examined the performance of aphasic patients, arguing that the combinatorial syntactic and semantic functions of language are cortically dissociable.

Similar to the vision network discussed earlier, language processing was claimed to involve functionally and anatomically distinct parallel dorsal and ventral pathways, which ground syntactic and semantic processes (Poeppel & Hickok, 2004). The dorsal pathway projects from the posterior part of the temporal lobe to premotor cortices and down to the inferior part of the prefrontal cortex via the arcuate and superior longitudinal fascicle (traditionally the major language pathway) whereas the ventral pathway connects the middle temporal lobe and the ventrolateral prefrontal cortex via the extreme capsule and unciate fasciculus (Figure 3.3-1). Recent studies using diffusion tensor imaging (DTI) offered similar claims where the dorsal pathway is associated with syntactic processing as well as articulatory sensory-motor mapping whereas the ventral pathway is associated with semantic processing (Catani, Jones, & Ffytche, 2005; Saur et al., 2008; S. M. Wilson et al., 2011).

Among the neural substrates related to this dual processing of the language network, the left pars triangularis (BA 45) and the left frontal operculum (BA 44), both of which are classified as classic Broca's area, are of our particular interest. While Friederici (2009) recently associated BA 44 and BA 45 with the dorsal and ventral pathway of the language network, the distinctive involvement of BA 44 and BA 45 in linguistic processing has been suggested by a number of researchers (e.g. Francisco Aboitiz & García, 2009; Amunts, Schleicher, Ditterich, & Zilles, 2003; Horwitz et al., 2003). Especially, BA 44 has been claimed to support the processing of hierarchically organized syntactic structures whereas BA 45 has been suggested to subserve controlled semantic processes (Friederici, 2002; Friederici, Opitz, & von Cramon, 2000; Friederici, Rüschemeyer, Hahne, & Fiebach, 2003). Similarly, the distinctive roles of BA 44 and BA 45 were suggested in terms of processing close class words (syntactic) and open class words (semantics) in a model of sentence comprehension (Dominey, Hoen, & Inui, 2006; Dominey, et al., 2009). A few studies on DTI also identified the differences in connectivity between these areas (Anwander, Tittgemeyer, von Cramon, Friederici, & Knösche, 2007; Frey, Campbell, Pike, & Petrides, 2008).

Therefore, BA 44 and BA 45 seem to have exclusive roles in language processing, each of which is limited to syntax and semantics, respectively. In fact, evidence suggests that BA 44 support a highly specific role in syntactic processing, especially

the process for handling complex syntactic structure. Stromwold and colleagues (1996) reported that BA 44 was activated more when judging semantic plausibility of syntactically complex sentences, such as center-embedded relative clause sentences, than less complex ones. Grodzinsky (2000) claimed that BA 44 supports the computation of the relation between transformationally moved phrasal constituents and their extraction (e.g. *Mary liked <u>which man</u> → <u>Which man</u> did Mary like?*). Similarly, Bornkessel and Schlesewsky (2006) argued that BA 44 is responsible for linearization of argument hierarchy during thematic role assignment. Moreover, the experiment result with deaf users of American Sign Language suggests that BA 44 involves in syntactic processing regardless of anatomy of the language articulators (Corina et al., 1999).

On the other hand, BA 45 was suggested to be involved in semantic processes. Hagoort and colleagues (2004) argued that left inferior frontal areas in the vicinity of BA 45 and BA 47 (the pars orbitalis) are involved in the integration of meaning and world knowledge during sentence interpretation. Gold and colleagues (2006) claimed that BA 45/47 support strategic retrieval of semantic representations in the lexical-semantic processing system. Moreover, the results of dynamic causal modeling (DCM) on verbal fluency tasks implied that BA 45 supports word retrieval processes whereas BA 44 is involved in processing phonological information during word generation (Heim, Eickhoff, & Amunts, 2009). Note that BA 44 and BA 47 are also suggested as the core areas of the semantics network proposed in the current work (Section 3.1).

Although we so far used the term "syntax" to address any set of generalized compositional rules in linguistic structures, caution is needed in relating BA 44 to syntactic processes in general. Rather, BA 44 seems to play only a partial role in syntactic processing, such as handling of intra-sentential dependency relation (Grodzinsky, 2000). In fact, it was demonstrated that grammatical ability is still retained to some extent in Broca's aphasics, who were reported to be able to interpret Japanese passive sentences with the canonical predicate argument structure (Hagiwara, 1993). Similarly, it was reported that lesions to Broca's area were not significantly influential in the Curtis-Yamada Comprehensive Language Evaluation (CYCLE)(Dronkers, Wilkins, Jr., Redfern, & Jaeger, 2004). According to Tettamanti and colleagues (2009), moreover, the left BA 44 was activated only during the acquisition of "non-rigid" syntax (distances of words are specified in terms of relative positions) as opposed to "rigid" syntax (a certain word must occur at a fixed distance). Thus, in linguistic processing there seems to be different degrees of grammatical complexity, from highly abstract global configurations to relatively fixed local structures based on transitional probabilities, both of which are grounded in possibly distinctive neural circuitries.

Interestingly, the latter type of grammar was suggested to be subserved by the "middle pathway", which is similar to the ventral pathway but ending in BA 45 (Friederici, 2006), implying a strong connection of simple grammars to the lexical-semantic processing system. This in fact blends a part of syntactic processing with lexical processing, suggesting that linguistic constructions vary at different levels of grammatical abstraction instead of being clearly separated into syntactic structures and the lexicons. The linguistic framework of the current work (i.e. Construction Grammar; Section 4.3) is coherent with this unified view of syntax and lexicon.

Another implication of BA 44 being involved in a very specific function in syntactic processing (e.g. manipulating transformational structures; Grodzinsky, 2000) is that linguistic abilities are widespread in brain regions rather than grounded in a small area solely dedicated for syntax. As discussed at the beginning of this section, left hemispheric perisylvian regions, such as inferior frontal regions around Broca's area, anterior, middle and superior temporal areas, and the inferior parietal

lobule, were all claimed to be responsible for linguistic processes, forming the language network.

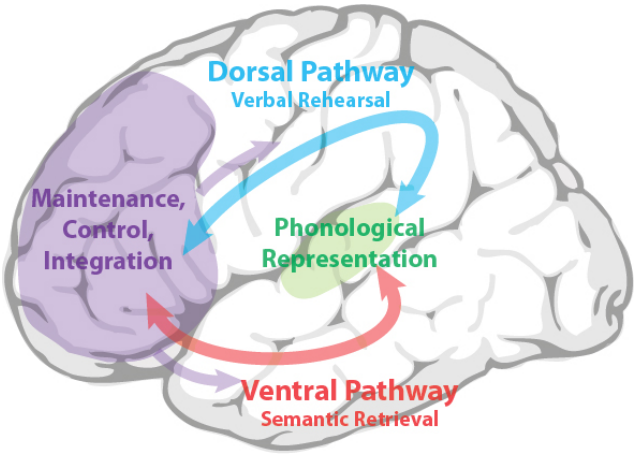


Figure 3.3-1: A schematized view of the language network forming a linguistic WM. The dorsal pathway (blue arrow), connecting posterior temporal areas with BA 44 of the inferior frontal gyrus (IFG) through the inferior parietal lobule, provides the rehearsal system of linguistic WM whereas the ventral pathway (red arrow), connecting middle and superior temporal regions with BA 45/47 of the IFG, provides the semantic retrieval system. These two distinct pathways run parallel while the DLPFC as well as the IFG performs executive and integration processes (purple), forming a WM for linguistic processes.

Consistent with the other networks addressed in the current work (Section 3.1 and Section 3.2), we view the language network as a type of WM dedicated to linguistic processes, which is illustrated in Figure 3.3-1. Although varied in subtle details, researchers generally agreed that the human language circuitries form a verbal WM which roughly consists of: (1) a verbal rehearsal component mediated by the left BA 44 and posterior temporal and inferior parietal areas, (2) a semantics retrieval component supported by the left BA 45 and middle and superior temporal regions, and (3) an executive component posited around the areas of the DLPFC (BA 46/9) (Francisco Aboitiz, Aboitiz, & García, 2010; Buchsbaum, Olsen, Koch, & Berman, 2005; Fiebach, Schlesewsky, & Friederici, 2001; Fiebach, Schlesewsky, Lohmann, von Cramon, & Friederici, 2005; Hickok & Poeppel, 2007; Poeppel & Hickok, 2004; Smith & Jonides, 1999; Smith, Jonides, Marshuetz, & Koeppe, 1998; Ye & Zhou, 2009). The first two components are grounded in the neural substrates for the dorsal and ventral pathways, respectively (Buchsbaum, et al., 2005; Hickok & Poeppel, 2007; Poeppel & Hickok, 2004), complying with the dissociative view of language processing. Moreover, the neural structures that ground the second component are also part of the earlier-proposed WM network of semantics and concepts (Section 3.1), implying the integrative nature of linguistic process. The PFC is also proposed to play executive functions in the semantics WM network, with more emphasis given to the ventrolateral portion of the PFC (VLPFC). In the current case, the dorsolateral portion (DLPFC) tends to get more attention as its involvement in mediating the overall language process, such as selective attention or task management, has been emphasized (e.g. Ye & Zhou, 2009). However, a recent view also emphasizes the importance of the inferior part of the PFC in linguistic processes. According to this view, linguistic operations take place in parallel at the semantic, syntactic, and phonological levels through the dorsal and ventral pathways, and the left inferior PFC around BA 45/47 plays an important role in integrating the information of those different levels (Baggio & Hagoort, 2011; Hagoort, 2005; Hagoort, et al., 2004).

Given that the language network stretches over various types of cortical areas, the functions of these so-called "language areas" may not necessarily be language-specific but rather shared with other cognitive abilities as well. Evidence suggests that the left inferior frontal lobe (BA 44/45), which plays a crucial role in syntactic processing, also participates in nonlinguistic functions such as visuo-motor and audio-motor coordination (R.-A. Müller & Basho, 2004). As the functional connection between Broca's area (especially BA 44/45) and manual actions has long been speculated (e.g. Michael A. Arbib, 2006), BA 44 was claimed to be responsible for object manipulation and the relevant movement control, supported by the observed activation during manipulation of various complex meaningless objects (Binkofski et al., 1999) and during imagery of forelimb movement (Binkofski et al., 2000). In fact, the suggestion is that the role of BA 44 is to manipulate any type of abstract sequential structures, which is manifested as handling syntactic structures in language domain (Hoen, Pachot-Clouard, Segebarth, & Dominey, 2006). Incongruities in musical harmony were reported to elicit activities in BA 44 (Maess, Koelsch, Gunter, & Friederici, 2001), supporting its involvement in abstract sequence handling.

A drastic view, moreover, sees the language network as a type of a "functional web" where widely distributed cortical areas in different modalities are linked together to produce correlated activity for processing the given linguistic information (Pulvermüller, 2001; Pulvermüller & Fadiga, 2010) – e.g. areas for word processing form a functional web with motor and premotor areas for the processing of specific kinds of words semantically related to arm or leg actions (Pulvermüller, Hauk, Nikulin, & Ilmoniemi, 2005). Thus, the areas of the language network may also contribute to other functions serving different purposes in different situations as the interaction of these areas is uniquely tied to the nature of materials and tasks employed (Kaan & Swaab, 2002). A number of neurophysiological studies reported cases where tasks in other domains utilize neural resources for linguistic tasks, supporting this view. For instance, the anterior and posterior portion of the left superior temporal gyrus, which are known to be involved in processing auditory speech, was also reported to be associated with processing of nonverbal auditory inputs, such as environmental sounds, pitch changes, or unfamiliar melodies (Price, Thierry, & Griffiths, 2005; Saygin, Dick, Wilson, Dronkers, & Bates, 2003).

As addressed in more detail in following Section 3.4, therefore, our view is that the language network, which we establish in the current work as a linguistic WM system, is grounded in a more large-scale WM network. In fact, it has been previously proposed that linguistic WM is embedded in a general WM network which interfaces incoming sensory information with the appropriate temporal organization of behavior and motor sequences (Francisco Aboitiz, García, Brunetti, & Bosman, 2006; Francisco Aboitiz & García V., 1997). In this particular view, Wernicke's area is originated as a converging place for multi-modal concept associations as well as phonological correlation while Broca's area is developed to be involved in performing complex vocalizations, resulting in the phonological-rehearsal system of linguistic WM. The projection between those two regions (through the arcuate fasciculus) is found to be absent or weak in nonhuman primates, suggesting the evolutionary specialization of the human brain for language (Rilling et al., 2008). Moreover, evidence suggests that this rehearsal system may be mere an emergent outcome of verbal procedures rather than a single established component operating as a continuous loop (Chein & Fiez, 2001; Jacquemot & Scott, 2006).

3.4. Integrative Working Memory Network

Given that the theme of the current work is to address some of the mechanisms involved in scene perception and verbal

description, the interplay between the vision and language system is of our particular interest. Previously, we addressed three sets of cortical structures that lay ground for the related processes. Those structures are the networks of semantics (Section 3.1), visual perception (Section 3.2), and language process (Section 3.3). These neural circuitries are proposed as working memory (WM) systems that are utilized for planning appropriate behaviors of an organism for the sensory input perceived from various modalities by developing and processing abstract concepts.

Although these WM systems have been discussed in a somewhat separate manner, our primary effort here in this section is to establish an integrative neural framework that addresses all of these systems. This integrative framework, which encompasses the cortical structures relevant to visual perception and verbal expression, provides a "workspace" for forming and maintaining abstract visuo-linguistic representations that are required to perform a scene description task. We call this integrative network the *Visuo-Linguistic Working Memory* (VLWM) since such an integrative network functions as a WM system, in which the interaction between the language and the vision system happens through constructing and manipulating visuo-linguistic representations. These representations are abstracted by the formal framework of the earlier-introduced SemRep (see Section 2.3 for detail), and we emphasize its particular role as an interface between the visual and linguistic aspect of the scene description process.

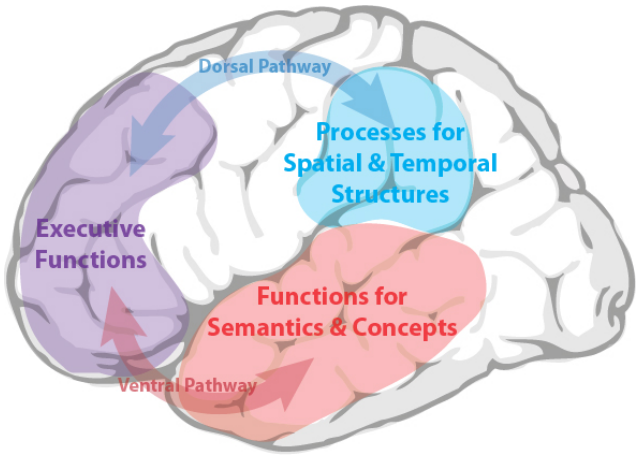


Figure 3.4-1: An illustration of the suggested integrative network of WM. The executive functions are centered in prefrontal areas (purple) while the processes for abstract structures and semantics and concepts are grounded in posterior temporal and inferior parietal areas (blue) and mid/superior temporal areas (red), respectively. WM processes for vision and language are incorporated within this general framework, resulting in the VLWM.

Before we move on, it is worth noting that our proposal of this integrative framework of language and vision, the VLWM, is based on the premise that there are a few components shared among the WM systems of semantics, language, and vision. As we will discuss in detail later, our claim is that the mechanisms of these WM systems can be explained in terms of a few common components described as follows (illustrated in Figure 3.4-1):

(1) The manipulation processes for *spatial and temporal structures*, suggested to be established around posterior temporal and inferior parietal areas, and BA 44,

(2) The accessing functions for *semantics and concepts* that are grounded in the mid/superior temporal cortex,

(3) The *executive functions* centered in prefrontal areas, including PFC, BA 45, BA 47, BA 46 and BA 9.

Especially the component (1) and (2) correspond to the dual processing mechanism previously mentioned in the vision and the language network – (1) is associated with the dorsal stream while (2) is associated with the ventral stream. Furthermore, the maintenance and the integration process, which are linked to the component (3), are believed to be established in the prefrontal cortex where both of the streams conjoin – the DLPFC has been suggested to be the end place for the dorsal stream whereas the VLPFC is for the ventral stream. The semantics, the language, and the vision networks addressed earlier all share the same structure to a certain degree, providing the foundation for a "general" structure of WM systems, in which the PFC plays the central executive role. In fact, a study that compared the activation patterns associated with tests of working memory, semantic memory, and episodic memory also reported that certain PFC regions, such as the left DLPFC and the left VLPFC, are commonly engaged across different memory tasks, indicating a general component in memory architecture (Nyberg et al., 2003).

Thus, the semantics, language, and vision systems tightly interplay with each other within a general WM framework manifested as the VLWM while SemRep is created and updated within the VLWM as the shared representations between those systems. This leaves us to question how the neural structures subserving the components discussed above are coordinated and how the information flow within those structures is handled in terms of SemRep.

Given that SemRep is an abstract form of a semantic representation, it is obvious that the component (2) is incorporated within the SemRep (see Section 2.4 for relevant details). The network of semantics and concept is fairly well established in terms of the ventral stream in the vision and the language system with many overlapping neural circuitries. As discussed earlier in Section 2.5, however, an important aspect of SemRep, which makes it distinguished from other semantic representations, is its emphasis on the "indexing" mechanism. SemRep contains not only the semantic meanings of the perceived objects but also specifies a set of parameters for guiding perceptual systems (i.e. in our case, the visual system). These parameters are supported by the dorsal stream of the WM network as opposed to the conceptual meaning or the identity of the perceived object, which is supported by the ventral stream.

While only the visual aspect of the indexing mechanism is emphasized in Section 2.5, however, one should note that this indexing mechanism is not necessarily limited to visual. Rather, it is supported by the circuitries for more general structures while the guiding parameters can be associated with positions within an abstract coordinate space, which goes beyond a spatial extent. These "indexes" can guide the executive processes to particular points of interest in the dimensional structure of the given task – e.g. in the vision case, they act as locational tags attached to particular objects under the current visual attention whereas in the language case, they are temporal (or sequential) markers attached to certain linguistic (syntactic or lexical) components relevant to the current discourse.

Thus, within our framework of the VLWM, this indexing mechanism incorporated in the SemRep is established by the neural circuitries supporting structural manipulation during visual and linguistic processes, which is addressed above by the component (1). So, the question is whether there is such a cortical network that supports manipulating abstract spatial-sequential structures and how the functions of the network are associated with the processes grounded in the dorsal stream. Rather than directly tackling the problem, however, we start with the component (3), the central executive function, and its neural foundations. Analysis on the nature of the central executive function (especially in terms of WM) and its tight

connection to attention will provide us with an explanation on how the processes of the vision and the language system are coordinated and performed within the integrative framework of WM and SemRep.

To begin with, the term "central executive function" has been used to refer to a domain-general factor of intelligence involved in WM-related tasks. McEvoy and her colleagues (1998) reported that the measurement of the subjects' event potentials in spatial and verbal versions of a visual n-back WM task showed two distinctive patterns, one of which varied with the type of information while the other was affected by the amount of information retained. Practice was reported to increase the response of the former type while the latter remained unaffected. Based on the findings, they claimed WM tasks draw upon task-general processes, which do not become more efficient with practice but are affected by the global attentional demand of the task. Similarly, confirmatory factor analyses and structural equation models indicated that WM tasks involve executive-attention processes that drive the broad predictive utility of WM span measures as well as domain-specific storage and rehearsal processes that relate more strongly to domain-specific aspects of complex cognition (Kane et al., 2004). In fact, many of currently accepted WM models conform with the idea of a central-executive component which is separated from domain-specific storage components – e.g. verbal rehearsal loop, or visuospatial sketchpad (Baddeley, 2003; Kane & Engle, 2002).

Among the WM models that agree with the idea of a separated executive component, moreover, some also suggested domain-specificity in their WM structure where the interrelationships between the components differ with the domains of tasks, leading to compositional asymmetry – the executive functioning and the storage component of the WM in the verbal domain were clearly separable whereas a much less clear distinction was implicated in the visuospatial domain (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001; Park, Lautenschlager, Hedden, Davidson, & Smith, 2002).

As briefly mentioned above in the component (3), many of WM studies that provided description on the central executive function (e.g. Baddeley, 2003; Kane & Engle, 2002) have pointed out the PFC as the neural ground for the domain-general executive component, or at least (possibly distributed) domain-general processes (Stuss, 2011). Supporting evidence comes from a number of earlier WM studies on primates. For example, in a study where monkeys had to remember a tone of a certain pitch with the color associated with it, the same prefrontal cells that responded to a particular tone were reported to respond selectively to the associated color, implying that the PFC represents behaviorally meaningful cross-modal associations (Fuster, Bodner, & Kroger, 2000). Similarly, the prefrontal cells of monkeys performing a delayed matching to sample task were reported to respond selectively to particular samples while their activation was maintained throughout the trial even when there was a disruption by intervening stimuli, suggesting that the PFC takes a role in executive functioning (Miller, et al., 1996).

Similar to the primate case, a number of WM studies on human subjects also reported congruent results. For example, Ranganath and colleagues (2004) reported that the dorsolateral PFC showed maintenance-related activity which was modulated by memory load regardless of the type of information in a delayed recognition task where subjects were required to encode a sequence of visual images (faces and scenes) and rehearse one of them after a delay period. The activity pattern of the PFC was significantly different from the activity patterns of other relevant areas, such as the fusiform face area (FFA) and the parahippocampal place area (PPA), which exhibited greater encoding- and maintenance-related activity when their favored stimulus (faces for FFA, and scenes for PPA) was presented. Based on the findings, they claimed that visual WM

encoding and maintenance processes are modulated by the prefrontal activity, suggesting the PFC's involvement in the executive process during a WM task. Moreover, Osaka and colleagues (2003) also claimed the PFC as a general neural basis for the central function in WM tasks. Despite differences in the task modality, increased activation in the PFC – especially, the connected network between the anterior cingulate cortex (ACC) and the left inferior frontal gyrus (IFG) – was observed during both reading span test and listening span test.

One interesting aspect regarding the PFC's executive role in WM tasks is that functional segregation of the lateral PFC has been suggested by a number of studies. These studies posit that the function of the dorsolateral part of the PFC (DLPFC) and the function of the ventrolateral part of the PFC (VLPFC) are distinguishable according to the type of process that must be performed in the task – the VLPFC mediates simpler maintenance processes such as the encoding and retrieval of information in a storage buffer while the DLPFC supports strategic memory organization processes involved in the monitoring and manipulation of information in WM (D'Esposito, Postle, Ballard, & Lease, 1999; Owen, Evans, & Petrides, 1996; Petrides, 2000; Rypma, Berger, & D'Esposito, 2002). Even though some studies attribute the segregation of the PFC to the type of information being processed rather than the type of process performed (e.g. Levy & Goldman-Rakic, 2000), researchers generally agree that distinct areas within the PFC perform different aspects of the executive process for WM tasks.

As mentioned earlier, our current focus is to locate the cortical network for manipulating abstract spatial and sequential structures and to investigate its functional implication in the integration of the vision and the language WM network. In fact, the evidence reviewed so far suggests that the DLPFC is such a network, and we are particularly interested in how it contributes to bridging between the vision and the language system.

In relation to this, it should be worth noting that there is a significant anatomical and functional overlap between the executive function and attentional processes. Especially, the DLPFC has significant anatomical and functional association with the dorsal stream processes that are established around the dorsolateral frontal and parietal network. Given the significance of this network in guiding eye movements and attention deployment (Curtis, 2006; Curtis & D'Esposito, 2003; Dominey & Arbib, 1992), the involvement of this region in the executive function in WM suggests a tight connection between attention mechanisms and executive functions in WM, especially in the visual domain. In fact, prefrontal areas, such as DLPFC, VLPFC, pre-SMA, have been reported to show sustained activity for attention orientation and short-term memory tasks (Nobre et al., 2004; Offen, et al., 2010). Similarly, Belopolsky and Theeuwes (2009) suggested that the neural system guiding eye movements is involved in coding and maintaining spatial WM. Curtis and his colleagues (2005) also claimed that the network of the FEF, the dorsolateral PFC, and the PPC support spatial WM by sustaining covert attention at a particular location.

Evidence suggests that the involvement of attentional mechanisms is not limited to spatial WM, but may also be extended to other types of WM, such as the WM of perceived visual objects. For example, Desimone and Duncan (1995) highlighted the similarity of top-down mechanisms in both object and spatial selection, and later Desimone (1998) reported that biasing of IT neurons in a WM task was remarkably similar to the biasing effects on the extrastriate cortex during visual search and spatially directed attention. Both of the studies pointed out the PFC as a main source of top-down feedback. Kastner and Ungerleider (2000) also emphasized the similarity between the biasing signals during object and spatial WM tasks in terms of their patterns and originating sources. The network formed between the PFC and the IT cortex has been

suggested to be an object memory circuit where bias from the PFC works in favor of behaviorally relevant stimuli (Desimone, 1998; Miller & Desimone, 1994; Miller, et al., 1996; Ranganath, et al., 2004). Note that the very same dorsal stream network, stretching from the PPC to the DLPFC, is also highlighted in the earlier-proposed WM network of semantics and concepts (Section 3.1) as to be responsible for manipulating associations among semantic items activated in WM.

Moreover, some studies suggested a possible overlap in the role of attention between the visual and the linguistic WM system. It has been claimed that the mechanisms of spatial attention are recruited in a rehearsal-like function to maintain information active in spatial WM, which is similar to the subvocal rehearsal within the phonological loop of language WM (Awh & Jonides, 2001; Postle, Idzikowski, Sala, Logie, & Baddeley, 2006).

Therefore, our view here is that attention control mechanisms have a particular role in the executive functions of WM, being involved in the organization and the sustainment of information in WM, possibly regardless of its modality. This view is in line with certain models of WM, which emphasize attention as the core component of the executive function. A good example is the model proposed by Cowan (1997, 1999), in which he claimed attention as the major component of WM processes, playing a central role in the retrieval of different types of items and their maintenance. These models are based on the observation that there is a huge anatomical and functional overlap between the mechanisms for attentional control and the central executive functions. In fact, a number of researchers have suggested that the function of the PFC in WM systems extends beyond a simple maintenance process to include aspects of attention, such as the inhibitory control or the goal-directed selection of information (Lebedev, et al., 2004; Miller & Cohen, 2001).

Similarly, Awh and colleagues (2006) emphasized attention's role in the executive control and highlighted its participation in the active manipulation and updating of the contents of WM, while comparing selective attention to a "gatekeeper" to WM. However, they also emphasized the multi-faceted nature of attentional processes involved in WM, arguing that selective attention should be defined as multiple stages of processing including both early sensory and post-perceptual processes rather than a single procedural component. A line of studies have proposed a similar view, claiming that attention is not a single mental construct that processes each item one by one, but rather, coherent "attention" develops as different systems compete and converge to work on related cognitive content (Desimone & Duncan, 1995; Duncan, 2006). Moreover, Postle (2006) claimed that even WM itself is an emerged property from the capability of representing many different kinds of information, controlled by behavioral goal- and task-related attentional biases.

One of the implications of these claims is that attention may function as a general cognitive mechanism rather than a specialized process limited to a certain domain, such as vision. The representations of the objects and events of many different cognitive contexts are recruited in a WM process through biased competition and selection mechanisms, which are collectively manifested as attention. The information of different modalities and domains is organized and integrated in terms of selective attention, thus allowing attention to serve a generalized executive function in WM. This is in accordance with the previously proposed claim in which attention is treated as the major component of WM processes, contributing to the retrieval and maintenance of different types of items.

This view is further supported by the results from a volume of behavioral studies that have indicated a tight link between attention and WM processes. The main argument of those studies is that the maintenance of information in WM involves covert shifts of attention, which are established generally as a "rehearsal-like" function similar to the subvocal articulation

within the phonological loop for maintaining verbal/phonological information (e.g. Awh & Jonides, 2001; Postle, et al., 2006). In fact, people appeared to have a limited attention capacity which is only large enough for just one "object" in WM at any one time (Garavan, 1998), suggesting that WM may recruit mechanisms for attention shifts for maintaining multiple items. Especially, a number of studies reported that the suppression or interference of attention rehearsal mechanisms impaired the performance of memory recall, supporting the claim – e.g. inducing saccades to irrelevant targets impaired the performance for visuo-spatial items, or rhythmic finger tapping had a detrimental effect on recalling a sequential order (Belopolsky & Theeuwes, 2009; Henson, Hartley, Burgess, Hitch, & Flude, 2003; Jones, Farrand, Stuart, & Morris, 1995; Smyth & Scholey, 1994; Tremblay, Saint-Aubin, & Jalbert, 2006; Zimmer, Speiser, & Seidler, 2003). Similarly, the manipulation of information during the retention interval of a WM task (e.g. rearranging the remembered material into an alphabetical order) was reported to cause an impairment on attentional processes, such as inattentional blindness (Fougnie & Marois, 2007).

Some studies have taken a step further, arguing for a more drastic view. In a recent paper by Iriki and Taoka (2012), for example, it was argued that crucial components of human intelligence, including language, would derive their character from the precursory spatial cognition process of the parietal cortex through the hominid evolution process. Similar studies have proposed that abstract mental operations, such as arithmetic calculations or sorting random sequences of digits, are supported by the parietal circuitry associated with spatial coding (Knops, Thirion, Hubbard, Michel, & Dehaene, 2009; Koenigs, Barbey, Postle, & Grafman, 2009; Noori & Itti, 2011). This particular view emphasizes the neural circuitries for attention shift as the core executive component for high-level cognitive tasks, claiming that the manipulation or rearrangement of information in WM is managed by the executive attention through a type of "spatial registry" system. The system for this spatial registry has been proposed to be grounded in parietal areas, especially in the superior parietal lobule (SPL). In fact, findings indicated that the attention rehearsal system for WM may be grounded in the prefrontal-parietal network that stretches from the DLPFC, which takes a role in directing top-down biases, to posterior parietal regions, which act as a modulator for those biases (Curtis & D'Esposito, 2003; Majerus et al., 2006).

Therefore, the suggestion is that higher-level cognitive processes that utilize WM networks are supported by attentional mechanisms, which are grounded in the parietal circuitries associated with spatial cognition. Especially within the current integrative framework between vision and language, our particular interest is whether linguistic processes also utilize these spatial circuitries that are generally utilized by visual processes. In fact, it has been suggested that the maintenance of verbal material utilizes the spatial rehearsal circuitry – top-down signals from the DLPFC select the relevant verbal representations in the inferior portion of the parietal cortex and Broca's area, enhancing those representations (Curtis & D'Esposito, 2003). Similarly, sequential/temporal order processing in language, such as syllable identification or word order comparison, has been suggested to be subserved by the left inferior parietal regions that act as a translator between auditory speech and articulatory maps in the dorsal stream (Hickok & Poeppel, 2007; Majerus, et al., 2006; Moser, Baker, Sanchez, Rorden, & Fridriksson, 2009).

Thus, the cortical structures for attention as well as spatial processing may play the role as the "neural bridge" between the vision and language system. Given with the current theme of the interplay between vision and language, this drives our attention back to the earlier question: whether there is a cortical network that supports forming and manipulating an abstract

visuo-linguistic representation – i.e. SemRep. In fact, the evidence we have so far addressed implicates that the prefrontal-parietal network for attentional processes might be such a cortical structure. While grounded within the dorsal processing pathway, this network performs the executive function in various cognitive processes by supporting the manipulation of abstract spatial and sequential structures.

Within our framework of the VLWM, this network plays a central role by supporting mechanisms related to SemRep, thus providing a shared workspace for visual and linguistic processes. The SemRep contains not only the semantic information of the entities but also the spatial (and temporal) "indices" that provide the constant referential frame where these entities can be located and maintained throughout the visuo-linguistic processes (see Section 2.5 for a relevant discussion). Linguistic processes, especially the ones that involve manipulating various abstract spatial and sequential structures from simple phrasal combinations to discourse-level narratives, may utilize this particular property of SemRep, whose relevant mechanisms are grounded within the prefrontal-parietal network for attentional processes.

In fact, some of recent studies provided particular cases where the language system exploits the spatial coordinate system. For example, Emmorey (2002) reported that signers assign a noun to a certain spot for later reference, whose distance approximately represents the proximity in space or the saliency in discourse of the noun. Emmorey also noted that signers use these referential indices in a similar way to how pronouns are used in spoken language, such as "this" and "that" in English. Similarly, Almor and colleagues (2007) reported that reading repeated names elicited more activation than pronouns in brain regions for spatial attention and perceptual integration (the middle and inferior temporal gyri and intraparietal sulcus), suggesting that the brain may rely on spatial processing to represent multiple linguistic references as if they were spatially distinct.

Our proposal of the integrative WM framework for vision and language extends along the same line, where attention mechanisms as well as their cortical structures establish the foundation for the process. Especially for a linguistic task involved with visual perception, such as scene description, attention plays a central role by closely administering the development and the access of the SemRep within the VLWM. Knott (2003) proposed a similar language model which links the sensorimotor sequence of attention to the scene (i.e. the scanpath) with the construction procedure of the syntactic tree for its description. He argued that the syntactic model of clause syntax can be mapped onto the sensorimotor model of action perception and execution, so that each node in the syntactic structure is characterized as a sensorimotor process while the hierarchical syntactic relationships between nodes indicate the sequencing relationships between these processes. However, his approach seems less pertinent than our approach of SemRep as the scene description is built on the simple structure of the eye movements rather than the state of the symbolic WM, thus limiting its capability in processing complex sentential structures such as recursion.

Chapter 4. Schema-based System of Utterance Generation

4.1. Construction Grammar[2]

The generative description as represented by Generative Grammar (Chomsky, 1965) seeks to explain language structure in terms of general syntactic rules, with any idiosyncratic properties derived from the lexicon. But how should linguistics treat idiomatic expressions like *kick the bucket*, *shoot the breeze*, *take the bull by the horns* or *pull strings*? Rather than taking their meanings as a supplement to general rules of the grammar, Fillmore, Kay, and O'Connor (1988) suggested that the tools they used in analyzing idioms could form the basis for a new model of grammatical organization, Construction Grammar, with constructions ranging from lexical items to idioms to rules of quite general applicability (Croft & Cruse, 2005). Thus, Constructionist approaches aim to account for the full range of facts about language, without assuming that a particular subset of the data is part of a privileged "core" (Goldberg, 2003).

There are a number of different versions of Construction Grammar (Croft, 2001; Fillmore, et al., 1988; Michaelis & Lambrecht, 1996), and they may differ in many ways, depending on the specific stances that they take. However, they all converge on a few basic tenets each of which represents a major divergence from the mainstream generative approach. Especially, a major tenet of most approaches to Construction Grammar is that "all" levels of description, including morphemes or words, idioms, partially lexically filled and fully abstract phrasal patterns, are understood in terms of *constructions* (Croft, 2001; Goldberg, 2003). 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. Thus, unlike the generative approach, Construction Grammar denies any strict distinction between the syntax and semantics – i.e. there is no principled divide between "lexicon" and "rules". Rather, it proposes a "syntax-lexicon continuum", which is intrinsically the same as the "lexicon-grammar continuum" of Cognitive Grammar (Langacker, 1987, 1991), while blurring out the distinction between simplex (lexicon) and complex (syntax) symbolic units – either kind may account as a construction.

However, the claim is not to discard the distinction (e.g. we can recognize words as distinct from phrases or sentences) but instead to suggest that we should extend the lexicon "upward" and syntax "downward" with no hard boundary but nevertheless with a different emphasis in each case. This is to acknowledge that we may treat a whole linguistic expression as the way we treat a lexeme while syntax provides a means to analyze it into finer components. This is similar to vision, where seeing the gist of a scene is akin to treating the scene as an object (c.f. see 2.6 for a detailed explication of this issue) while we still need other mechanisms to integrate perception of separate objects into the scene. We may understand some sentences with little or no syntactic processing while others require more subtle constituent analysis. For example, the exclamatory sentence "*I am sorry*", especially when spoken out after stepping on someone's foot, is never understood as "Noun Be-verb

[2] A number of sentences in this section are written based on the lecture slides presented at Conceptual Stucture, Discourse and Language (CSDL) Conference – Arbib, M. A. (2010). Construction Grammar: A tutorial.

Adjective" but rather as a whole as an apologetic expression, whereas the sentence "*the dress is blue*" may be. An important point is that even when a phrase has syntactic structure, we may remember it as a whole, with its associated meaning – just as we may remember words whether or not we know their etymology. The construction in Construction Grammar is proposed to capture such subtleties in holistic semantics and structural configurations in linguistic expressions.

Thus, the argument so far rejects a strong form of "compositionality", whose principle is that the meaning of a complex expression is fully determined by the meanings of its components and the way they are combined. Although a scene consists of a number of objects, a simple combination of those objects is not necessarily equal to what the scene represents. The emphasis is that words do contribute some, but not all the meaning, as evidenced in the following example[3]:

(1) ***Bill hit*** *the jackpot, the lucky bastard.*

(2) ***Bill hit*** *Mary, the poor waitress.*

The semantics of (1) and (2) are significantly different although they share almost the same surface structure – especially, the constituent *Bill hit* represents a very different meaning in each case. Thus, the "construction" refers to the more abstract structural template of grammatical features (the type) rather than to the specific complex constituent (the token) – e.g. sentence-level constructions may have their own schematic meanings, which are independent of those of the verbs and other constituents combined.

This is to be contrasted with approaches based on Generative Grammar, in which autonomous syntactic rules put words together in very general ways and without regard for the meaning of the result. The meanings of idiomatic expressions, such as "*he pulled strings to get the job*", cannot be predicted on the basis of its parts. Like what Generative Grammar argues for, if such "exceptional" expressions were fixed in form, it would be conceivable to add them to the lexicon. However, many of them also have grammatical structure. For example, the *X's way* construction, such as "*Bill kicked his way through the crowd*", consists of a particular syntactic structure, which is roughly "Subject Verb X's way Oblique". According to Kemmerer (2006), this syntactic structure is paired with a particular semantic structure, which roughly means "X makes progress along path Y by Verb-ing".

Goldberg (2006) argued that the well-known sentence "*Pat sneezed the foam off the cappuccino*" exemplifies the case more clearly. The meaning of the sentence cannot be understood solely by the typical meaning of the verb "sneeze", which only contains intransitive sense as used in the sentence "*Pat sneezed*". However, "sneeze" in this case appears in a "syntactic" construction like the former case, with the syntactic structure of "Subject Verb Object Oblique", whose meaning roughly corresponds to "X causes Y to move from Z by Verb-ing". If we regard the sense of "cause to move something by sneezing" as a property of the verb "sneeze", then it should have at least two different senses, one for the typical intransitive sense and the other specialized for this case. However, it seems quite implausible, at least in English. Thus, the burden of explanation should not be entirely placed on the verb itself (Evans & Green, 2006) – it should also be on the syntactic construction and

[3] This is a modified version of the original example in the lecture slides presented at Conceptual Structure, Discourse and Language (CSDL) Conference – Arbib, M. A. (2010). Construction Grammar: A tutorial:

Bill hit *Mary, the cad.*

Bill hit *the jackpot, the lucky bastard.*

the related semantic knowledge (e.g. sneezing carries plosive force).

In fact, it has been claimed that verb meaning should also be interpreted in terms of the highly schematic sense which goes beyond the concrete meaning unique to particular verbs (Iwata, 2005; Rappaport Hovav & Levin, 1998), and indeed action verbs often occur in constructions describing a wide range of action events. For example, even though *kick* is usually considered to be a prototypical transitive verb, it occurs in at least nine distinct constructions, each of which describes a different "scene" involving a different number/type of arguments (Goldberg, 1995).

Table 4.1-1: Examples of constructions with different argument structures (adapted from Table 10.1 of Kemmerer, 2006).

Example Sentence	Construction	Form	Meaning
Bill kicked the ball.	Transitive	Subject Verb Object	X acts on Y
Bill kicked the ball into the lake.	Caused motion	Subject Verb Object Oblique	X causes Y to move along path Z
Bill kicked at the ball.	Conative	Subject Verb $\text{Oblique}_{\text{at}}$	X attempts to contact Y
Bill kicked Bob the ball.	Ditransitive	Subject Verb Object_1 Object_2	X causes Y to receive Z
Bill kicked Bob black and blue.	Resultative	Subject Verb Object Complement	X causes Y to become Z
Bill kicked Bob in the knee.	Possessor ascension	Subject Verb Object $\text{Oblique}_{\text{in/on}}$	X contacts Y in/on body-part Z
Bill kicked his foot against the chair.	Contact *against*	Subject Verb Object $\text{Oblique}_{\text{against}}$	X causes Y to contact Z
Bill kicked his way through the crowd.	X's *way*	Subject Verb X's way Oblique	X makes progress by performing action
Horses kick.	Habitual	Subject Verb	X performs action habitually

Thus, a purely "bottom-up" or lexically driven model of grammar fails to provide the whole picture, but rather sentence- or clause-level constructions themselves carry meaning, to some extent independently of the words in the sentence. Constructions are themselves theoretical primitives rather than "taxonomic epiphenomena (Chomsky, 1991)" as they should be understood in terms of the "whole-self" that goes beyond the surface form of the expression, such as grammatical category or a particular word order.

4.2. Support for Construction Grammar

In Radical Construction Grammar, Croft (2001) argued that we cannot assume the existence of language-independent universal categories (e.g. noun, verb, adjective) as providing the grounding for grammar, but instead, categories are derived from the construction(s) in which they appear. He pointed out that the criteria used for grammatical categories in some languages are either completely absent in others or are employed in ways that seem bizarre for those brought up on English. For example, Vietnamese lacks all inflection, and Makah has inflection but employs it in a surprising manner – e.g. it applies

aspect and mood markers not only to words for actions that are translated into English as verbs, but also to words for things and properties that are translated into English as nouns and adjectives (Kemmerer & Eggleston, 2010).

Although we may cross-linguistically identify prototypical nouns as specifying objects and prototypical verbs as specifying actions, it appears that human languages contain an open-ended spectrum of historically shaped, constructionally based, hierarchically organized, and distributionally learned grammatical categories. In other words, languages evolved culturally as the collectivity of many properties through a process of "tinkering" that added, combined and modified constructions.

This idea is in sheer contrast to the notion of Universal Grammar (Chomsky, 1965), which basically asserts that there are "innate" (not learned through "tinkering") properties that all possible natural human languages have. According one version of UG, a particular language is a collection of structures with properties resulting from the interaction of fixed principles with parameters set one way or another in the child's environment (Chomsky, 1991).

However, studies on language acquisition demonstrate that all linguistic knowledge is constructed by learning within a language community as the child's constructions (including lexicon) are shaped through experience to better approximate usage within the community. Constructions are "incrementally" constructed during the language learning period as more substantial and complex patterns of construction (e.g. embedded clauses) emerge from a source of relatively simple patterns of construction (Diessel & Tomasello, 2001; Israel, Johnson, & Brooks, 2000).

For example, the child may first acquire what the adult perceives as two-word utterances as a holophrase "*want-milk*", which the child later develops into a more general construction "want X", in which X can be replaced by the name of any "wantable thing" (Michael A. Arbib, Conklin, & Hill, 1987; Hill, 1983). With further experience, the child will develop more subtle constructions, such as "Verb Object", with word classes like "verb" or "noun", which are defined by their abstract thematic roles rather than particular meanings in a particular situation, such as "wanting" and "wantable thing". Similarly, Matthei (1982) tested understanding of the phrase "*second green ball*", and found young children interpreted the phrase "as the ball which is second and green", as opposed to the adult interpretation which is "the second of all the green balls". The children's misinterpretation was attributed to children's tendency to use of "flat" structures, which would be replaced by hierarchical ones later in language acquisition. Again, the suggestion is that at a different developmental stage, the structural complexity of learned constructions varies, even for simple phrasal structures, refuting the innateness of grammatical knowledge.

Furthermore, Verhagen (2010) argued that the capacity for dealing with center-embedding patterns (i.e. recursion) is not built into humans (with the maximum of only two or three embedded structures; Lewis, 1996), but rather it has a cultural foundation, especially from the development of writing systems which might have facilitated the recursive use of grammatical patterns by providing an extension of memory. This suggests that even a highly elaborate linguistic capability, such as recursion, might not be "innate" in human but rather acquired later as necessary.

Kemmerer (2000, 2006) was among the first to make explicit the relevance of Construction Grammar to neurolinguistics by presenting the major semantic properties of action verbs and argument structure constructions. In a lesion study, Kemmerer (2000) reported a double dissociation in the impairment of the pure verb semantics and the constructional meaning; two patients performed well, but one patient performed poorly, on a word-picture matching test that required them

to discriminate between verb triplets that differed only in semantic features attributable to the lexical meaning of the verbs (e.g. *spill-pour-sprinkle*), whereas the first two patients performed poorly, but the other patient performed well, on a second test that required them to judge the grammaticality of sentences containing the very same verbs (e.g. "*Sam spilled beer on his pants*" vs. "*Sam spilled his pants with beer**"). Similarly, Kemmerer (2003) also reported experiment results where two patients demonstrated a dissociation in judging subtle aspects of the pure verb meaning and the grammar-relevant constructional meaning in the English body-part possessor ascension construction – e.g. "*Sam hit Bill on the arm*" vs. "*Sam broke Bill on the arm**". Patients passed a verb-meaning test (again, word-picture matching tasks) while they failed a grammaticality judgment test on constructions with the same verbs. Based on these findings, Kemmerer concluded that the neural substrates of constructional meanings are separate from those for verb meanings, arguing for the existence of constructional knowledge in the brain.

Recently, Allen and colleagues (2012) demonstrated that the dative construction (e.g. "*Sally gave the book to Joe*") and the ditransitive construction (e.g. "*Sally gave Joe a book*") were distinguished (through Region-of-interest MVPA analyses) by the activation patterns within areas of left BA 47 and anterior BA 22. Since the distinction does not rely on traditional language areas, such as the left BA 44/45 or the left posterior BA 22, they argued that the distinction was not based on syntactic differences, especially involving the word "to". Rather, they claimed that particular grammatical constructions that shared the same content words, propositional meaning, and degree of surface complexity can be distinguished based on neural correlates. Such separability appears to lend support to neural substrates for distinctive construction types.

Moreover, it has been claimed that grammatical constructions affect mental representations (i.e. meaning) of described events. Bergen and Wheeler (2010) reported that progressive sentences (e.g. "*John is closing the drawer*") generated a significant Action-sentence Compatibility Effect (ACE; Glenberg & Kaschak, 2002) – facilitatory priming of manual actions by sentences denoting similar actions (e.g. "*close the drawer*" implies action away from the body) – while perfect sentences (e.g. "*John has closed the drawer*") did not. These two types of sentences were identical in every way except for their aspect as they shared the same content words that are arranged in the same order, thus allowing Bergen and Wheeler to conclude that constructional meaning contributes to the higher-order semantics of a described event, which goes beyond what is provided from the semantics of content words.

4.3. Template Construction Grammar

We now present our own version of Construction Grammar, *Template Construction Grammar* (TCG) (for the earlier version, see Michael A. Arbib & Lee, 2007, 2008), as the linguistic framework. Since the main purpose of the current work is to provide a computational model that explains certain phenomena related to describing a visual scene, TCG is proposed to capture the dynamics of scene perception and generation of verbal expression. As addressed in detail in Chapter 5, findings from studies on scene perception and speech production as well as the results from eye-tracking experiments we conducted indicated that the process of scene description requires constant interaction between the vision and language system so that the tight coordination between the two systems becomes crucial. During the scene description process in TCG, a number of constructions are activated simultaneously as "schema instances", and they cooperate and compete in order to produce a verbal description of a scene. This process is proposed to be highly dynamic in that:

(1) The SemRep is constantly updated based on the current foveation, which both depends on and helps drive interpretation by the vision system,

(2) the verbal structure for the utterance is formed in an incremental way as more constructions are attached, building atop those constructions which link directly to the current SemRep, and

(3) the verbal output need not be a fully formed sentence and can be generated at any moment.

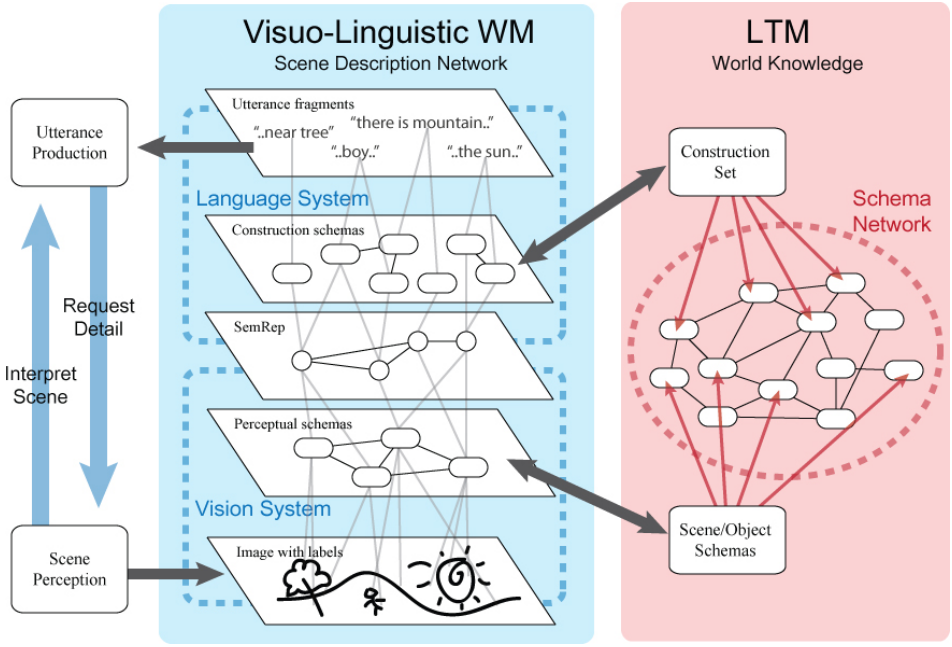


Figure 4.3-1: A highly dynamic process of scene description in which the vision system and the language system are concurrently running and constantly interacting. The Visuo-Linguistic Working Memory (VLWM) provides a workspace where various types of schemas are collaborating (according to the competition and cooperation paradigm) to produce a verbal description of a scene. The vision and language system are continuously interacting with each other as the vision system concurrently interprets the scene and updates the SemRep while the language system occasionally generate requests for more details from the vision system.

The key idea of the proposed process is that basically two systems, the vision and the language system, are running in parallel while the SemRep is acting as a "middle-ground" representation of the two. The language system applies constructions on a SemRep and reads off the formed sentences or sentence fragments that result while the vision system concurrently interprets the scene and updates the SemRep. The processes of these two systems are tightly correlated such that the vision system provides the interpretation of the current scene on which the language system works on, driving the scene description process, while the language system sometimes generates requests for more details to the vision system, biasing the scene perception process.

Both the vision system and the language system are proposed here to be schema-based: the vision system (presumably a system similar to the VISIONS) generates a SemRep, which is an abstraction of perceptual schemas, and the language system (currently TCG) deploys constructions, which are implemented as schemas, to produce verbal expressions. During the process, both of the systems develop a shared working memory space, the Visuo-Linguistic Working Memory (VLWM) (see

Section 3.4 for relevant neurophysiological accounts), in which the SemRep is constantly updated to represent the "current" interpretation of the scene, constructions are applied on the SemRep updated at the moment, and a partial or full syntactic structure that are formed so far are read out. The scene description process of TCG is both: (1) "incremental" since new constructions are constantly applied according to the updates on the SemRep, and (2) "hierarchical" since constructions may be applied to bind other constructions already attached to the SemRep, forming a more complex syntactic structure. During the process, the SemRep acts a type of an "anchoring" system by which different types of conceptual and cognitive representations are organized, accessed, and associated with each other.

Thus, the production process of TCG (see Section 4.4 for more detailed accounts) is performed in such a highly interactive and dynamic manner that new construction instances are constantly invoked as the SemRep keeps getting updated while some of the instances are connecting to other instances, and others are competing with each other. As a whole, the system reaches to the solution through the competitive and cooperative interactions between construction instances.

Given that the purpose of the present work is to propose a schema-based computational model for scene perception and description production, selecting Construction Grammar appears suitable in a few aspects.

Firstly, constructions, which are basically defined as pairs of form and meaning, provide a detailed account of the mapping between a particular syntactic pattern (as low as phonological ordering) and a particular semantic pattern (as high as event structure). Constructions of various levels of grammatical complexity and semantic coverage can be manifested as a set of encapsulated processing units linking the semantics of a perceived scene (provided as perceptual schemas) to the production of the corresponding utterance (performed via motor schemas). Thus, from our schema theory point of view (Section 2.1), constructions in our framework can be regarded as a type of "coordinated control program (Michael A. Arbib, 1981)", or more simply coordinating schemas, since each of them acts as such a processing unit mediating the coupling of perceptual and motor schemas.

Secondly, in the current work, we provide the SemRep as a unique approach for representing the semantics of a perceived visual scene. The semantics of the scene, such as objects, actions, and their relations, are represented in a form of graph structure (i.e. nodes and edges), and transforming it to a verbal expression requires a particular type of linguistic framework. We regard Construction Grammar as a suitable framework since the construction as a whole accounts for the abstract structural template of semantics and syntax, which can be directly mapped to a certain subpart of SemRep and the corresponding linguistic structure. A construction, especially a complex one, such as a sentence-level construction, may go beyond the surface form of the expression as it means more than a simple combination of the meanings of its constituents (see Table 4.1-1 for example constructions associated with particular event themes exceeding the lexical meaning of the verb). In an analogy with vision, a construction may be matched with a subscene (Section 2.7) as the "meaning" of the construction represents the gist of the subscene while the "form" represents the event structure by which separate items of the subscene are integrated into a coherent scene. Considering the tight correlation between vision and language emphasized in the current work, Construction Grammar may provide the most straightforward conceptual foundation for translating visually perceived events (i.e. SemRep) into linguistic expressions.

However, one should note that TCG is not necessarily to be bound to the Construction Grammar framework as the notion "construction" is not particular for Construction Grammar. In fact, although TCG adopts two major policies of conventional

Construction Grammar (i.e. each construction specifies the mapping between form and meaning, and the systematic combination of constructions yields the whole grammatical structure), TCG also differs from other Construction Grammar approaches in some important aspects. As indicated earlier, constructions are implemented as "schemas" which cooperate and compete to reach an overall resolution on verbal description production, and the semantics of constructions are defined as subgraphs of the SemRep, which are directly mapped to the representation of a perceived scene. Most of all, a crucial feature of TCG is its capability to form complex grammatical structures, especially recursive combinations among constructions, such as the embedment of relative clauses.

When Hill (Michael A. Arbib, et al., 1987; Hill, 1983) developed a computational language model based on her observation on the child's acquisition of a general construction "want X" developed from an initial holophrase-like expression "*want-milk*", she used the term "template" for addressing such a general construction, highlighting its role of providing a standardized pattern that defines the categorical constraints on its components (e.g. X in "want X" template is replaced by any "wantable thing"). We adopt Hill's conception of the term template in naming our approach as "Template" Construction Grammar, focusing on the role of constructions as semantic and syntactic templates. TCG's ability to form complex grammatical structures is accomplished by encoding categorical constraints in both the semantics and syntax within the template of a construction, especially a construction of a complex structure.

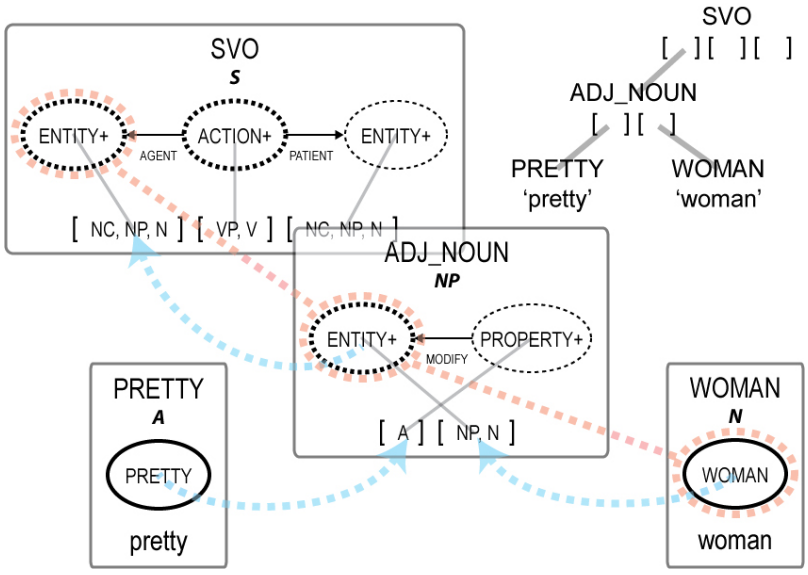


Figure 4.3-2: An illustration of how a grammatical structure is built in TCG. When the semantics of the "head" components (rather than the conventional sense of "head", which is a word) of a construction (represented as nodes with thick lines) is matched with the associated semantics of the slot that the construction fills in (e.g. ENTITY and WOMAN in combination of ADJ_NOUN and WOMAN construction) and the class of the construction is matched with one of the classes specified in the slot (e.g. N of WOMAN construction and [NP, N] of the second slot of ADJ_NOUN construction), syntactic combinations between constructions (represented by the blue dashed-line) are made. The head components of the combining constructions act as the "pivot" in the combination of the three constructions (WOMAN, ADJ_NOUN, and SVO) as they play a role as the representative components of the constructions that fill into the slot of other constructions (the ENTITY nodes are successively replaced by the WOMAN node through the syntactic linkage as represented by the red dashed-line) – e.g. in forming the phrase *pretty woman*, the head node WOMAN and the associated construction becomes the head of the

phrase, *woman*.

Such a complex construction is mostly distinguished by its usage of syntactic “slots”. A construction may contain one or more slots that can be filled by other constructions of less grammatical complexity – constructions without slots (similar to the WOMAN or PRETTY construction shown in Figure 4.3-2) just act as simple lexical items, such as words, that can “fill into” constructions with slots that represent relatively complex grammatical structures, such as phrases, clauses, or even sentences (similar to the SVO or ADJ_NOUN construction in Figure 4.3-2). In TCG, through this combinatorial mechanism, the syntactic hierarchy between constructions is established and complex sentential structures (even with recursion) are built. Such complex constructions provide abstract structural templates of the grammatical and semantic features of the language. Their slots, which constrain the semantic meaning as specified by the coupled SemRep elements and the syntactic category as specified by the marked classes, represent “generic features” that will be specified thereafter. For example, the concept ENTITY in SVO in Figure 4.3-2 limits the semantics of the “head” of the combining construction with the first slot of SVO to be an entity rather than an action or a property while the marked classes, NC, NP, and N, limit the syntactic category of the combining construction to be one of those classes that are equivalent to the conventional noun clause, noun phrase or noun. In combination with the specification of the head components of a construction, the slots allow the system to formulate grammatical structures with different types of constructions, eventually enabling the system to exhibit the full-fledged capability for dealing with recursive structures. Through the slots, constructions are combined with other constructions, thematic roles are assigned, and grammatical hierarchy is formed.

Such an approach seems quite similar to other theories of syntax that put emphasis on the syntactic head in building grammatical structure – i.e. the so-called “head-driven” approaches. According to those approaches, such as Head-driven Phrase Structure Grammar (HPSG) (Levine & Meurers, 2006; Pollard & Sag, 1994) or X-Bar theory (Chomsky, 1970; Jackendoff, 1977), phrasal structures are defined by the heads of the phrases (e.g. the noun for a noun phrase). These approaches are based on the idea that linguistic structure is organized around lexical entries, so it is necessary to highlight a certain lexical item to be the head (i.e. the dominant item) in order to build the syntactic hierarchy among them. Thus, the head is the key constituent to build the sentential hierarchy by bundling other lexical items together. However, although we generally agree on the importance of the head, the position that we take in TCG differs such that the significance of the head in TCG is not only syntactic but also semantic – a slot imposes constraints on the semantic meaning (through the concept of the associated SemRep elements) as well as the syntactic category (through the specified classes).

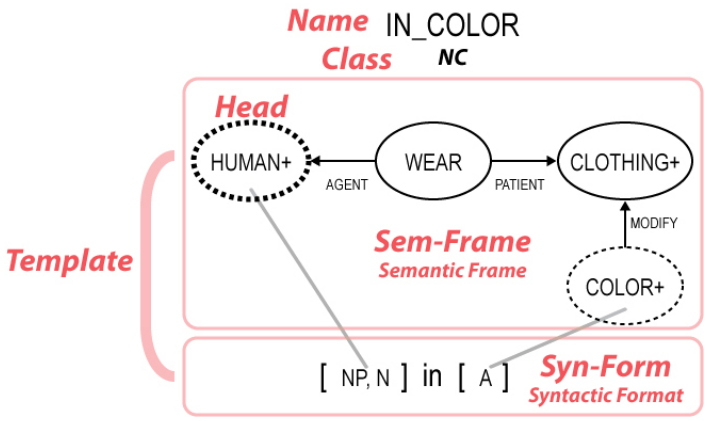


Figure 4.3-3: A schematic view of a typical construction in TCG. This particular construction, named IN_COLOR, corresponds to an idiomatic expression to describe a person wearing an outfit of a certain color (e.g. *man in black*, *woman in blue*, etc.). See text for more detail on each field.

As illustrated in Figure 4.3-3, a construction in TCG is defined by a triple (name, class, and template) with a few subfields where:

I. A construction is assigned with a *name*, which is not involved in the language process – it is only there for reference purposes.

II. *Class* specifies the "category" of the result of applying the construction, which could serve as a specifier in combination with other constructions. In general, class is not the same as the conventional syntactic category, such as noun or verb, since it represents a more-or-less subtle combination of the syntactic and semantic type of a construction.

III. *Template* defines the form-meaning pair of a construction, and it has two subcomponents, Sem-Frame and Syn-Form, that correspond to the meaning and form part of the construction, respectively.

IV. *Sem-Frame* (semantic frame) of template defines the meaning part of the construction. Its meaning is defined by the part of a SemRep graph that the construction will "cover" – the literal meaning of covering since TCG produces verbal description by basically having constructions cover subregions of a SemRep (see Section 4.4 for more detail). Some elements can be coupled with the syntactic slots defined in the Syn-Form, constraining the semantics of the head components of the constructions that fill in. Added to that, Sem-Frame also specifies the "head" components which act as the representative of the whole construction when forming hierarchy with other constructions.

V. *Syn-Form* (syntactic format) defines the form part of the construction. It consists of a series of phonetic notations, which represent words or morphemes, and *slots*, which specify the type of constructions that will be combined within (see Section 4.4 for more detail). A slot acts as a "generic" syntactic component, whose semantic meaning is constrained by the coupled SemRep element in the Sem-Frame, and whose syntactic category is constrained by the specified classes associated with the slot.

VI. A construction may be given a *preference value* which incorporates a various personal and linguistic factors (e.g. expression frequency, personal preference, or priming effects, etc.)

Consider the construction IN_COLOR illustrated in Figure 4.3-3. The class NC roughly corresponds to a noun clause in the conventional sense as the construction contains a subject-centered event description (i.e. noun-like) that fits into the syntactic structure of a clause. The head component (emphasized with the thick line), which is the HUMAN node in the Sem-Frame, specifies the representative meaning of the whole construction during syntactic combination (i.e. in the phrase *woman in blue*, the word *woman* is the head). Moreover, although an element defined in the Sem-Frame is generally a typical SemRep graph element and its formatting follows that of a conventional SemRep, it may have extra features than an ordinary SemRep element. The HUMAN node illustrates such an example – its concept is specified as *inclusive* (represented by a "+" sign after concept) while its covering region is specified as *shared* (represented by the dashed line). A concept specified as "inclusive" acts as a generic semantic type as HUMAN+ is judged to be matched with all subcategory level concepts, such as MAN, WOMAN, BOY, or VICTOR. Moreover, an element specified as "shared" can be overlapped with other elements (of other constructions) without conflict when covering a SemRep, allowing combination between constructions to happen at that overlapped area (see Section 4.4 for more detail). For this reason, all of the Sem-Frame elements coupled with slots are specified as shared. Furthermore, the first slot of IN_COLOR specifies NP and N as the possible classes of the constructions to be filled in, indicating that the slot can be replaced by only the constructions with the matching classes (i.e. constructions of either N or NP class). Combining with the coupled concept HUMAN+, the constraint imposed on the type of constructions that can fill in this particular slot is that the class of the construction should be N or NP while the concept of its head element should be defined as a subcategory-level concept of HUMAN, such as WOMAN. Through slots, a construction imposes semantic and syntactic constraints during grammatical formulation.

Thus, a construction acts as a syntactic template as well as a semantic template – the Syn-Form specifies the syntactic structure of the construction and constraints on its grammatical combination with other constructions while the Sem-Frame posits constraints on semantic categories that the construction should represent. The implication of this particular feature of TCG is that the application of constructions in TCG is intrinsically "bi-directional". Although only the production process is considered in the current work, the same construction set can be used for comprehension too. As opposed to the production procedure where the Sem-Frame initially acts as a template for selecting proper constructions by constraining the semantic categories and the topological structure of the SemRep graph, the Syn-Form becomes the template imposed on the verbal expressions provided as the input during the comprehension procedure. The selection of constructions is done by matching the phonetic information of the input words as well as their grammatical positions (e.g. sequential order in English) as constrained by plausible semantics.

Another particular feature to note in TCG is that the information encoded in a concept is considered as the "semantico-syntactic knowledge" of the associated entity (see Section 2.4 for the detailed exposition). It is basically a combination of syntax-oriented features and semantics-oriented features, which includes the *conceptual meaning* of the entity (i.e. the capitalized label attached to a SemRep element, such as WOMAN) as well as other semantic and syntactic properties, such as *gender*, *person*, *number*, *definiteness*, and *tense*. Although these properties are not represented in concepts in Figure 4.3-2 or Figure 4.3-3 (only the conceptual meanings are shown), they are also necessary in judging similarity between concepts during the process of TCG. The following provides some example concepts appeared in the figures above:

WOMAN concept (in Figure 4.3-2)

```
Meaning: WOMAN
Gender: female
Person: 3
Number: 1
Definiteness: no
Tense: unspecified
```

WEAR concept (in Figure 4.3-3)

```
Meaning: WEAR
Gender: unspecified
Person: 3
Number: unspecified
Definiteness: unspecified
Tense: present
```

HUMAN+ concept (in Figure 4.3-3)

```
Meaning: HUMAN+
Gender: unspecified
Person: unspecified
Number: unspecified
Definiteness: unspecified
Tense: unspecified
```

As mentioned earlier, the HUMAN+ represents an inclusive concept, which is designed to be used in a complex construction for providing a generic semantic type (only in a construction) as opposed to the regular concept (i.e. specific), which is to represent the conceptual meaning of an entity (both in a SemRep and a construction). Generally in TCG, inclusive concepts are in superordinate category levels whereas regular concepts are specified in basic category levels – they are claimed to be the most codable, most coded, and most necessary categories in language (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976).

Theoretically, each concept is considered to be an abstract representation of assemblages of perceptual schemas that deliver parameterized information for the properties and conceptual meaning while the entire conceptual knowledge is represented as a schema network where semantic processes (e.g. retrieval or comparison) are done through activations of schemas within the network. Thus, for example, judging semantic similarity between LAD and HUMAN is done through a chain of activations within the schema network, starting from both the LAD and HUMAN schema. Although the current version of TCG does not provide further details on such a network of schemas and the currently implemented version adopted (Section 4.6) a purely symbolic approach, we propose that the schema network of TCG should be implemented in a neurophysiologically plausible way, such as a connectionist network. In fact, Shastri and Ajjanagadde (1993) demonstrated a high-performance inference machine (performing a class of inferences with millions of facts and rules in a few hundred milliseconds) based on a connectionist network.

Furthermore, the Syn-Form of a construction in the current version of TCG is defined as a simple sequence of words and slots since the word order is the major characteristic of English syntactic structure. However, for other languages, the Syn-Form needs not be such a simple sequence but may be a much more complex structure. For example, in Korean or Japanese, the subject and object are specified by the grammatical particles while the order is relatively less important. In this case, the Syn-Form might include the information of the particles too. The order information should be represented in a less strict manner than English as the positions of the subject and the object in a typical Korean or Japanese sentence are interchangeable – sometimes, even the verb, which generally appears at the final position, can come to the head of the sentence without damaging grammatical and semantic integrity.

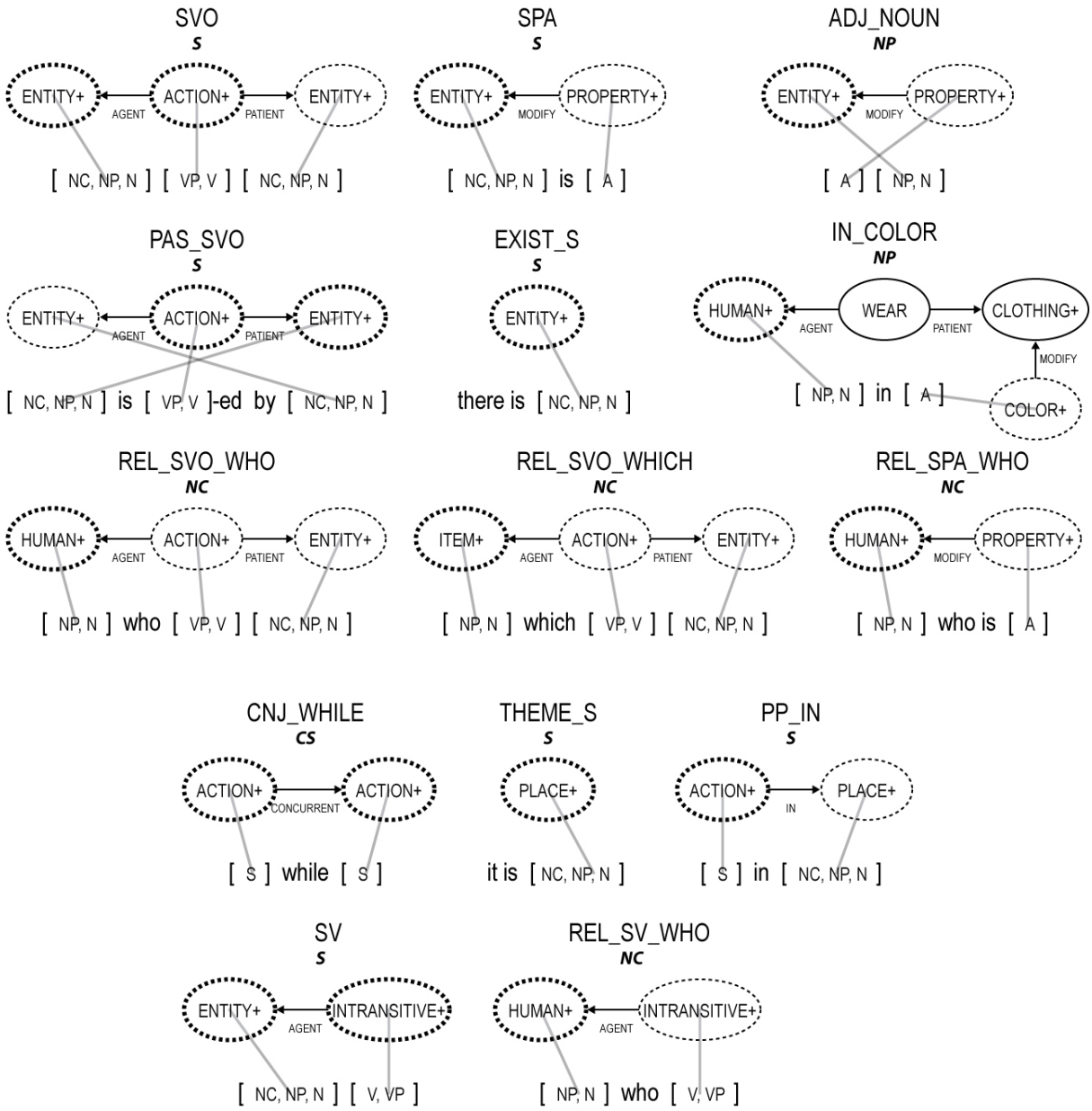


Figure 4.3-4: An example set of "complex" constructions that correspond to relatively complex structures, such as sentences or clauses. The syntactic structures are represented by the arrangement of slots and phonological notations in the Syn-Form and the coupled SemRep elements in the Sem-Frame. The head elements are depicted with a thick line. Note that some constructions have classes that also appear in one of their slots, whose coupled Sem-Frame elements are the same as their heads (e.g. ADJ_NOUN, IN_COLOR, CNJ_WHILE, and PP_IN). That feature allows those constructions to be recursively connected (e.g. ADJ_NOUN can fill in the second slot of another ADJ_NOUN).

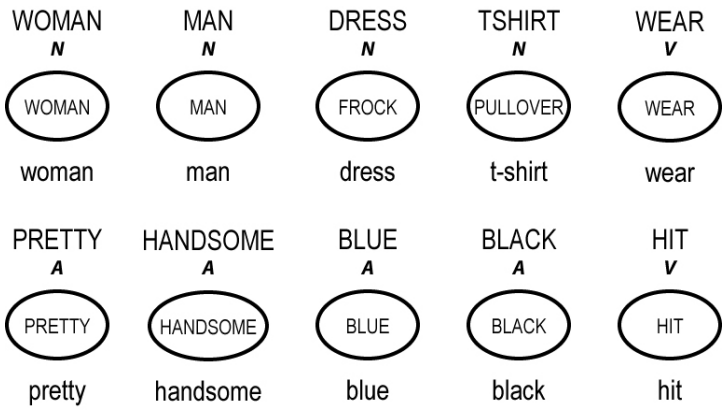


Figure 4.3-5: An example set of "simple" constructions that are mostly lexical items, such as words. They are generally represented as simple pairing of the semantic meaning (represented as a single node, which is as well the head element) and the corresponding phonetic expressions. Note that the concept (i.e. signified) is not the same as the word (i.e. signifier) – DRESS and TSHIRT construction illustrate such a difference.

Although Construction Grammar does not distinguish lexicon from grammar in principle, constructions defined in TCG can be broadly categorized into two types, based on the information that each construction encodes. Simple constructions (Figure 4.3-5) mostly encode lexical information, and they are the ones that generally associate words with concepts. On the other hand, complex constructions (Figure 4.3-4) encode more abstract syntactic information, and they generally correspond to sentential or phrasal structures. They may also add words that do not directly label elements of the corresponding portion of the SemRep, such as function words – e.g. EXIST_S in Figure 4.3-4 contains the words *there is*, which are not directly matched with any of the nodes in the Sem-Frame (the only node is matched with the slot). In his Grammatically Relevant Semantic Subsystem Hypothesis (GRSSH), Pinker (1989) argued that there is a distinction between two large-scale components of meaning: (1) a set of fairly abstract semantic features that are relevant to grammar insofar as they tend to be encoded by closed-class items as well as by morphosyntactic constructions; and (2) an open-ended set of fairly concrete semantic features that are not relevant to grammar but instead enable open-class items to express an unlimited variety of idiosyncratic concepts. Similarly, Talmy (2000) suggested that the cognitive representation provided by language can be divided into lexical and grammatical subsystems.

However, our position taken in TCG is separable from the conventional lexicon-grammar dichotomy even if the account on such a distinction appears to conform to the generative grammar point of view. We would argue that the distinction between the simple and complex construction does not originate from any innate judgment, but it is built quite arbitrarily depending on the empirical function of a construction shaped through linguistic experiences – i.e. it is tacit knowledge. For example, the IN_COLOR construction does not conform to categorization in the conventional sense as it contains both highly grammatical components (slots) and semantic components (specific semantic requirements for the noun and the specifier). Thus, we would rather relate the grouping of constructions (as denoted by their classes) shown in Figure 4.3-4 and Figure 4.3-5 to ontogenetic factors rather than to the conventional grammatical structures and categories. Croft (2001) also emphasized such empirical factors in defining construction categories – for him, the category of a word seems to be the set of slots it can fill across all constructions. Similarly, in his account of distributional analysis, Verhagen (2009) argued that words

and phrases do not form large, distributionally uniform classes (e.g. noun, verb, adjective, etc.) with which major classes are differentiated from subclasses, but rather word classes are language-specific and cannot be identified cross-linguistically on formal grammatical grounds. Thus, the classes of constructions and their resultant groupings in TCG need not follow the conventional sense of syntactic category while they might be fine-grained in some places and coarsely defined in other places.

For example, the ditransitive construction, whose schematic meaning is "X causes Y to receive Z", puts subtle restrictions on the acceptable verb categories – verbs of instantaneous causation of ballistic motion are acceptable (e.g. "*I kicked / tossed / rolled / bounced him the ball*"), but verbs of continuous causation of accompanied motion are not (e.g. "*I carried / hauled / lifted / dragged him the box**") (Pinker, 1989). However, the ditransitive construction is not sensitive to the more fine-grained contrasts between the verbs within each set, meaning that the restriction is construction-specific. Moreover, Pinker (1989) pointed out that some speakers find the sentences with verbs of accompanied motion to be acceptable, suggesting dialectal or idiolectal differences in dativizability.

Although the argument so far emphasizes the ad-hoc nature of word categories, it would also be nonsense to imagine that every word must have "yes-no" data for every slot of every construction. It seems more likely that we have a fairly large category and a branch of the category is learned where a particular word varies from the slot-set for that category, establishing an inheritance relationship, and enough encounters with the subcategory may define a new category. Class in TCG is proposed to capture such established categories.

4.4. Production Process of TCG

The production process of TCG consists of three subprocesses: the invocation, cooperation, and competition process. During the invocation process, the system invokes new construction instances based on the SemRep maintained in the VLWM. The Sem-Frame of a construction in the repertoire is matched with the currently updated part of the SemRep, and when matched, a schema instance of the construction is created and attached to the matching part of the SemRep. When invoked, a construction instance is attached to a certain region of the SemRep, and this region is considered to be "covered" by the instance. The goal of the system is to cover as much area of the SemRep without conflicts. These newly invoked instances are maintained within the VLWM until their activation values are faded below a certain level. Through the cooperation process, construction instances are combined with other construction instances when their syntactic and semantic structures are compatible, eventually forming a group of instances (i.e. forming a hypothesis). Construction instances in the same group strengthen each other's "suitability" for the solution, eventually allowing the larger group to have more chances (yet not always) to be chosen for the solution. The *suitability* of a construction instance represents how suitable the instance is to be included in the finally produced utterance. Construction instances, or groups of instances, also compete with each other if they are in conflict through the competition process. A conflict happens when a construction instance tries to cover the region which is already covered by other construction instances. When there is a conflict, the system tries to assess the suitability for the conflicting instances by measuring a number of factors, including the semantic coverage and closeness, the ease of production (mostly by the syllable length), personal preference, and other semantic and syntactic constraints. The suitability of an instance is assessed in terms of the whole group of connected instances. When competing, construction instances with lower suitability will lose the competition and lower their activation levels, eventually getting eliminated from

the VLWM.

Now we focus on a more detailed description of each of the invocation, cooperation and competition processes.

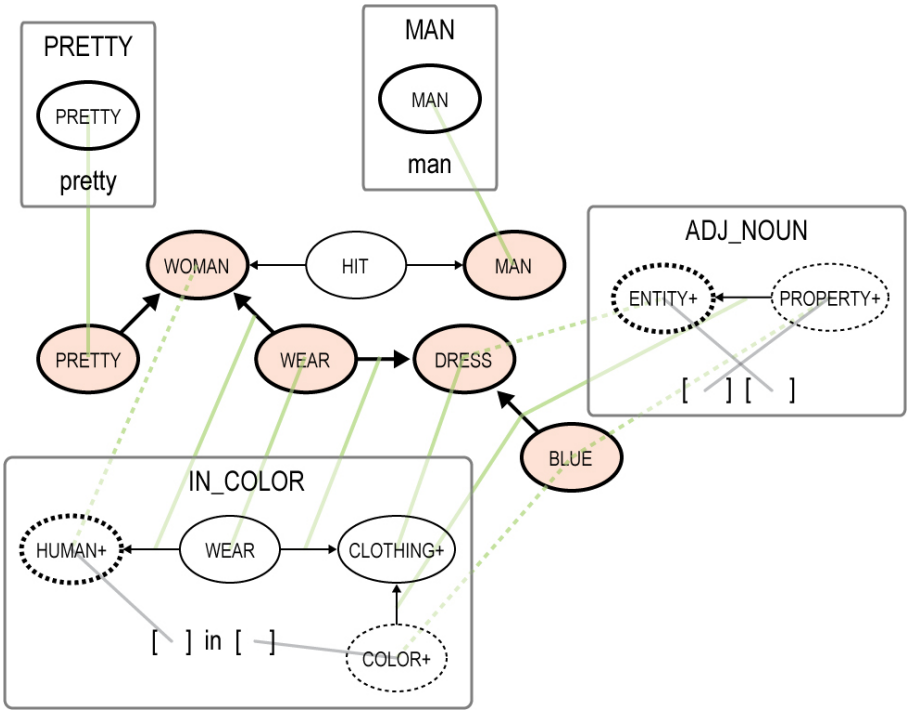


Figure 4.4-1: A number of construction instances are invoked over parts of a SemRep, covering the areas where they are invoked over (slightly colored red). These invoked instances are matched with their covering areas both “semantically” (i.e. matching concepts of graph elements) and “topographically” (i.e. matching graph structure). Note that some nodes are denoted as “shared” (dashed lines), and it means that their covering is not “exclusive”. Some details of constructions are omitted for clarity.

The invocation process mainly concerns the creation of new construction instances. The SemRep maintained within the VLWM is constantly updated as new information is perceived from the scene. The system matches the Sem-Frame of constructions with the SemRep, and when matched, instances of the matching constructions are created, or “invoked”. These newly created instances are attached to the matched regions of the SemRep, and they are called to “cover” those regions. Basically, the production process of TCG is to cover as much area of the SemRep as possible.

When a construction is being matched with a certain region of the SemRep, a few conditions are in consideration, which are outlined as follows:

- The topology of the Sem-Frame of a construction needs to be matched with the area of the SemRep that the construction is going to cover. Topology means the arrangement of nodes and relations.
- The concepts of all elements in the Sem-Frame of a construction need to be matched with the elements in the region of the SemRep. In order to be matched, the conceptual meanings of the concepts need to be “similar enough” (based on the activation of the schema network) and all of the subfields (e.g. gender, number, etc.) need to conform.

Newly invoked instances are assigned with a certain initial activation value, and they are maintained within the VLWM while being attached to their covering regions until the activation values drop below a certain level – construction instances

with activation values below that level are eliminated.

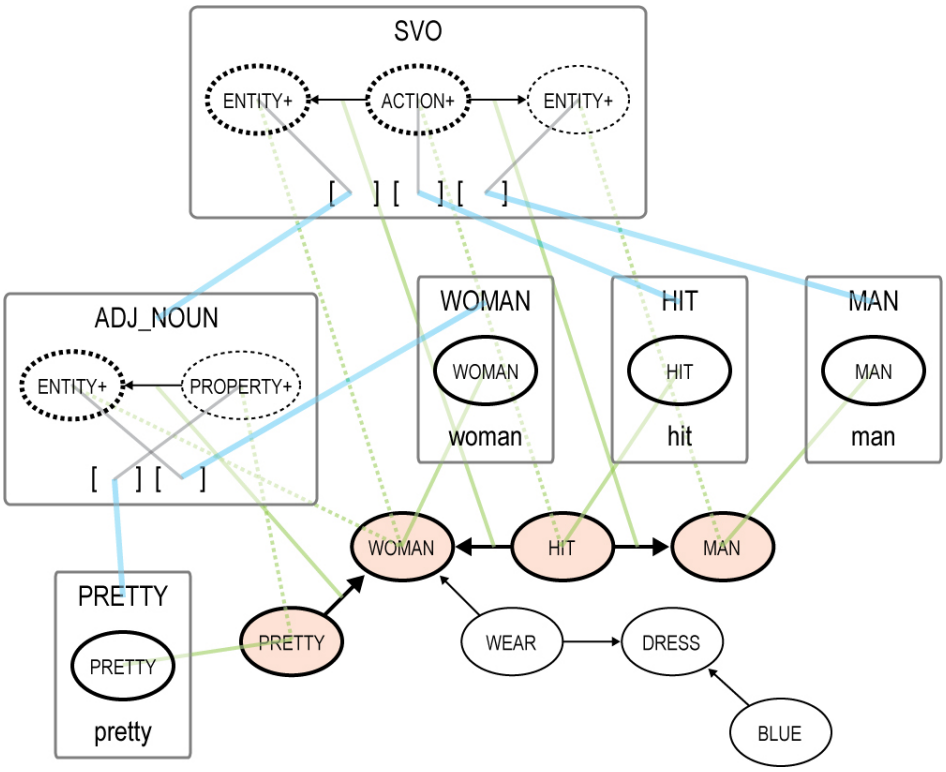


Figure 4.4-2: Invoked construction instances are combining into a group, forming a hierarchical structure, which as a whole represents a grammatical expression (in this case, the sentence "*pretty woman hit man*"). Constructions with the overlapping covering areas are combined when their heads and classes match with the requirements specified by the slots (although the classes are not shown in the figure). Note that some of Sem-Frame elements are mearked as shared, thus allowing combination can be made by avoiding conflict between constructions – e.g. in SVO, all nodes are denoted as shared, allowing ADJ_NOUN, HIT and MAN are slotted in without conflict.

Active construction instances can be combined with each other, forming a group of instances, which as a whole is a type of syntactic structure representing a verbal expression, such as a phrase or a sentence. The cooperation process in TCG concerns such a combination between instances since construction instances increase their suitability for the solution by combining with other construction instances, extending the covering area of the entire group (i.e. representing more semantics). Cooperation between construction instances happens when their covering areas overlap without any conflict, which is possible through the usage of the shared elements in the Sem-Frame. The shared elements are supposed to cover the regions of the SemRep in a nonexclusive manner. When the covering areas overlap, the system checks the class as well as the head of the construction instances to see if any of them can be inserted into others' slots.

More precisely, when a construction instance is combining with another construction instance, the following conditions should be met:

- A construction instance covers an area which overlaps with another construction instance's covering area. The overlapping areas do not conflict – i.e. the covering elements (nodes or relations) of either of the instances are defined as "shared" in the overlapping area.

- The class of one of the instances is matched with the class specified in the slot of another instance, which is associated with the Sem-Frame element(s) covering the overlapping area.
- The overlapping element(s) of the instance that fills in another instance's slot is specified as the "head".

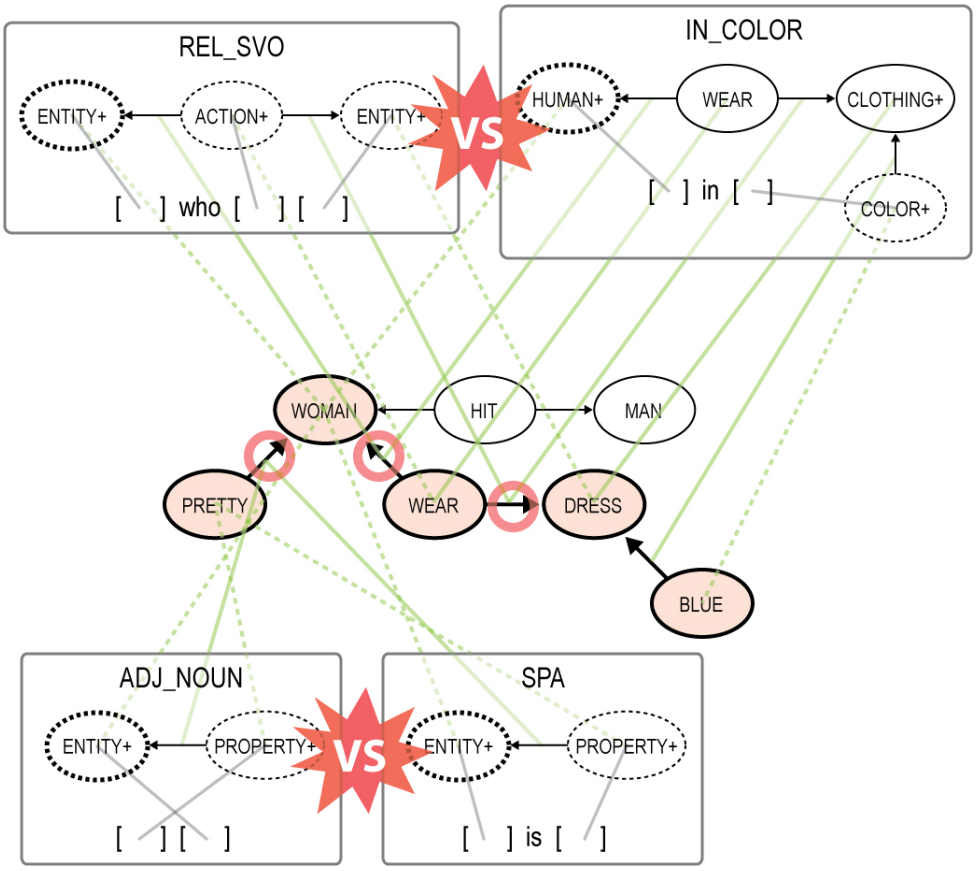


Figure 4.4-3: Two groups of construction instances (REL_SVO & INCOLOR and ADJ_NOUN & SPA) are competing as their covering areas are in conflict (the red circles). Conflict only happens for the regions covered by non-shared elements – areas covered by shared elements are not in conflict.

Construction instances compete with each other if their covering areas are in conflict. Conflict happens when a construction instance tries to cover the region that is already covered by another construction instance. The rationale of the competition process is to filter out construction instances representing "redundant" meanings (i.e. covering the same areas of the SemRep), and only the winning construction instances are selected for the final solution (i.e. the produced utterance).

When there is conflict between instances, the suitability of the conflicting instances is calculated by assessing various semantic and syntactic factors. When the suitability of an instance is being assessed, it is done for the whole group that the instance is connected (i.e. cooperation of instances) rather than for the particular instance. If the instance is connected to more than one group, the best suitability is taken.

The following provides specific measurement factors for assessing suitability:

- Among other measurement factors, the semantic similarity receives the most priority. It measures how much and close the semantic meaning of an instance represents the semantic meaning of the SemRep. It is basically a

combination of counting the number of covering SemRep elements and assessing the similarity of the element concepts.

- Yet with less priority, the expression preference is considered as well. As Flores d'Arcais (1975) indicated, people tend to show structural preference in producing utterance. Although not rigorously considered in the current version of TCG, a construction may be assigned with a preference value – e.g. the active voice (SVO) is more favorable than the passive voice (PAS_SVO), or a sentence-level construction is preferred over other constructions (as the system favors to produce a proper sentence), etc.
- The ease of production is also taken into account. The simplest way for measuring is to count the number of syllables that the construction, or the group of the constructions as a whole, would produce, and it may extend to address the ease of articulatory muscle movements as well.
- Other syntactic and semantic constraints, such as the priming effect or the task requirements, may be put in the decision.

When two constructions compete, the construction instance with a lower suitability value will lose the competition, and its activation level is decreased. An instance with an activation level lower than a certain threshold will be eliminated.

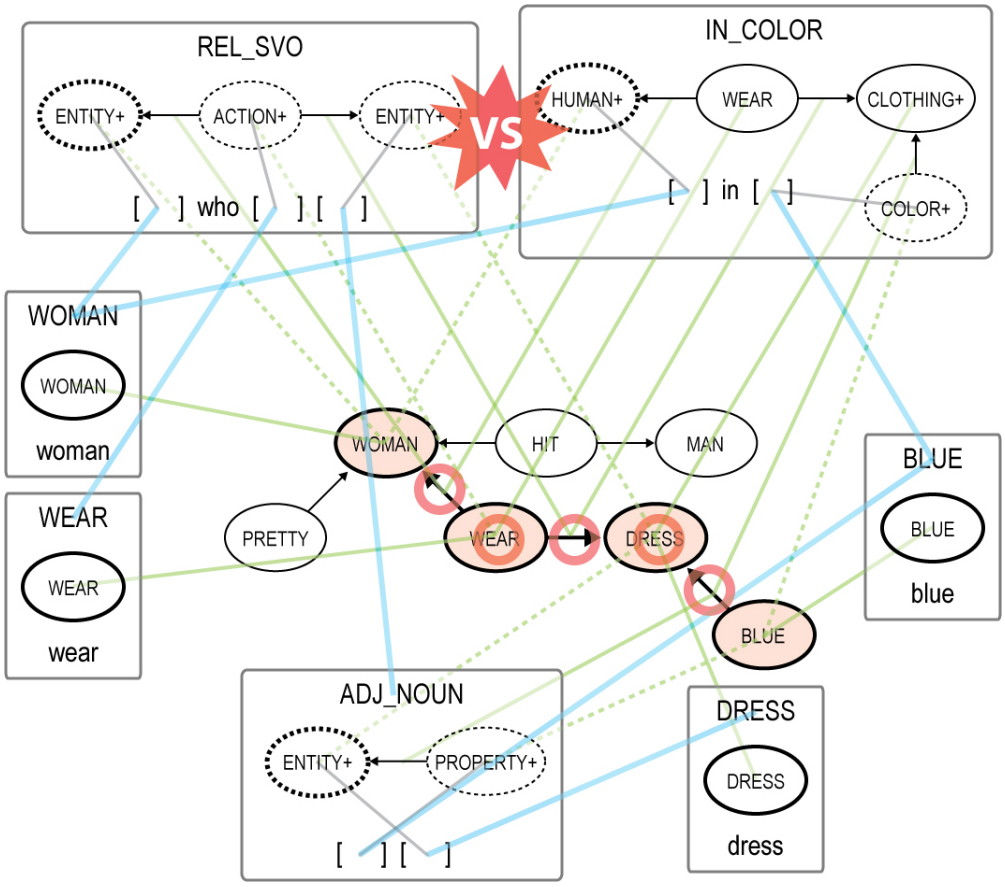
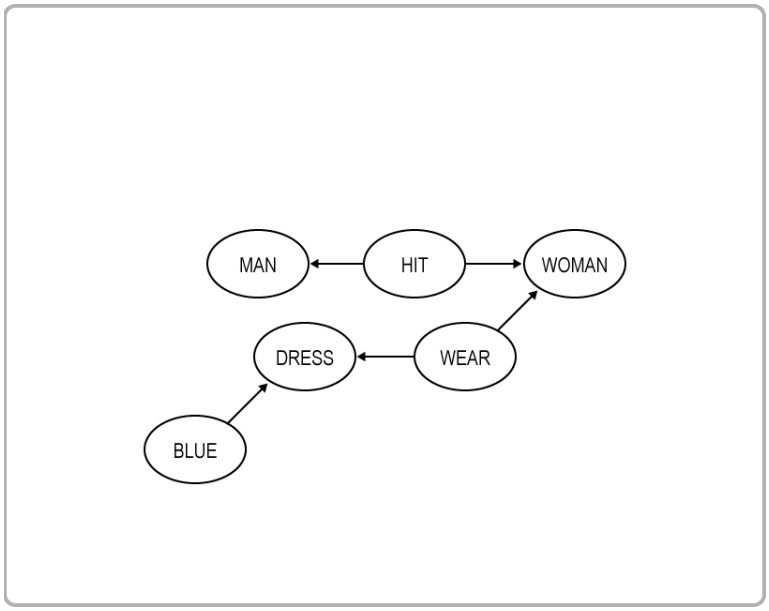


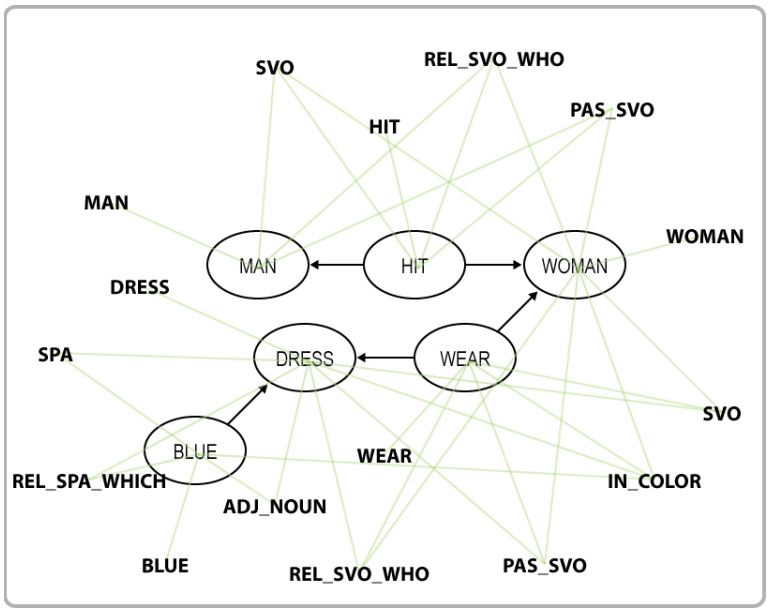
Figure 4.4-4: A snapshot of the production process of TCG, in which invoked construction instances are forming groups while some of them are competing at the same time. Through the competitive and cooperative interactions between construction instances, the system reaches to the solution – i.e. a verbal expression for describing a scene.

Although the invocation, cooperation and competition processes have been so far described in a separate manner, these processes are actually performed simultaneously during the scene description process. The vision system constantly updates the SemRep in the VLWM as more information is perceived from the scene while the language system continuously applies construction schemas on the update part of the SemRep. At the same time, some of construction instances cooperate, forming a group of instances, while the others compete or get eliminated from the VLWM. Some of the grouped construction instances are read out intermittently during the process when conditions are met. Thus, the production process described here is a dynamic blend of diverse traces of processes whose interactions are established through the shared working memory.

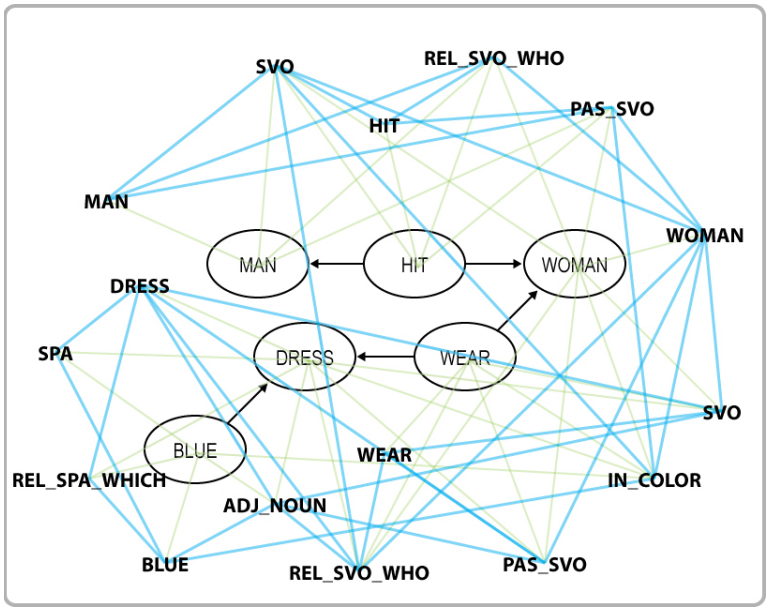
The following is a series of illustrations of an example simulation of the production procedure in TCG, based on the construction set depicted in Figure 4.3-4 and Figure 4.3-5. In real cases, the simulation output is much more dynamic, with constant SemRep updates and interleaving cooperation and competition processes between construction instances, but it should be worth considering a static (and unreal) case to better understand how the system works. Note that although the three processes of invocation, cooperation, and competition are represented in separate illustrations, they perform concurrently in real simulation.



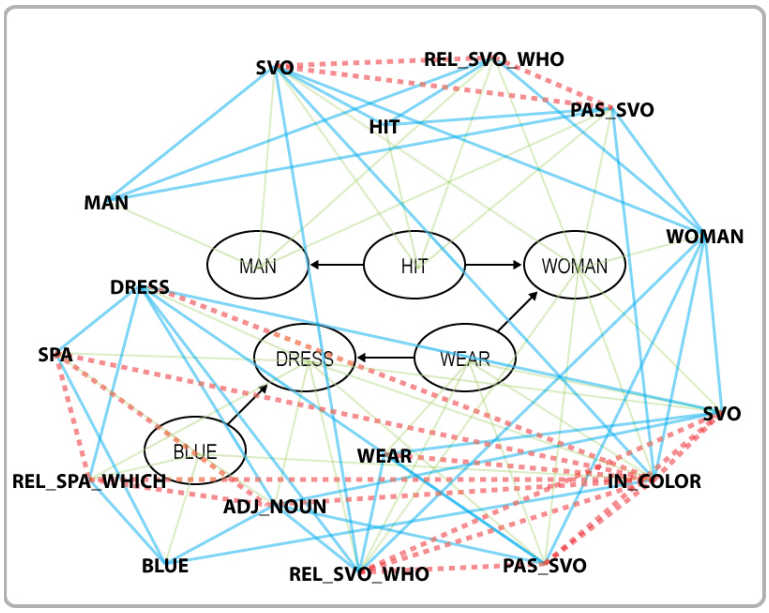
(1) A SemRep is provided from the vision system and stored in the VLWM. In fact, this is hardly the case that such a big SemRep like this has been provided at once since the vision system interprets the scene in an incremental manner, possibly in terms of subscenes (Section 2.7). The SemRep represents two events, *woman is hitting man* and *woman is wearing a blue dress*. Some details are omitted for clarity.



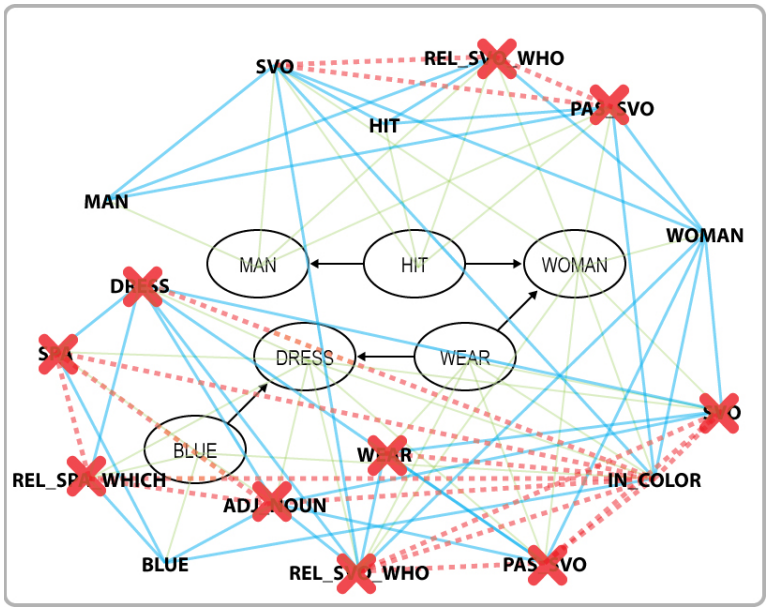
(2) Construction instances are invoked over the SemRep. Not all of invoked constructions are shown (e.g. EXIST_S constructions are supposed to be invoked over the WOMAN, MAN and DRESS nodes) and some details are omitted for clarity (e.g. the covering of a construction instance, which is shown as green lines, are shown only for nodes, and only the names of constructions are shown).



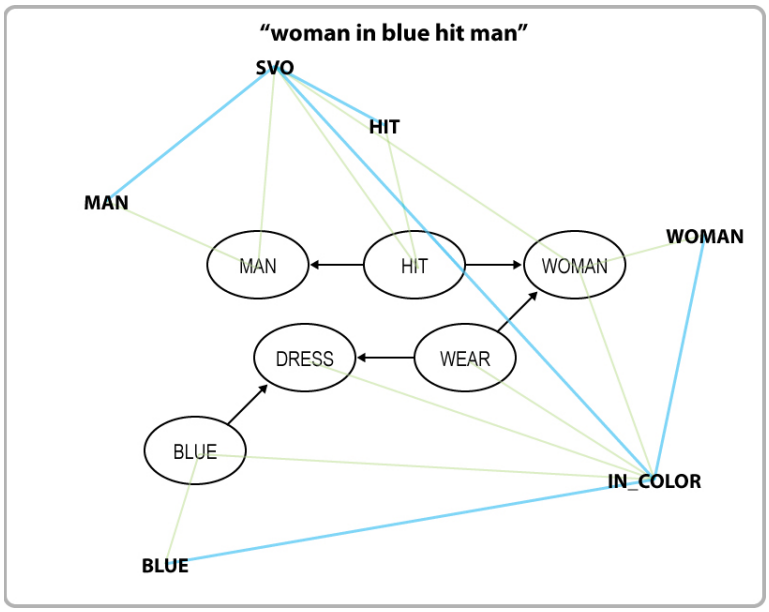
(3) Construction instances are combining each other when the conditions for cooperation are met (e.g. covering the same area, classes matched, etc.). Note that some instances (e.g. WOMAN) are combined with multiple groups of instances as all possible combinations are under consideration during the cooperation process. They represent partial solutions (i.e. hypotheses) that the system forms at the moment (e.g. *woman who wear dress hit man, woman in blue hit man, blue dress is worn by woman who hit man*, etc.).



(4) Since some of construction instances are covering the same area of the SemRep where their Sem-Frame elements are not marked as "shared" (e.g. SVO, REL_SVO_WHO and PAS_SVO all cover the nodes MAN, HIT, WOMAN and the relations between those nodes, which are not specified as shared), competition happens. Competition is done in terms of groups of instances rather than individual instances in conflict, meaning that the suitability values of those conflicting instances are assessed in terms of the cooperative structures that they belong to.



(5) After competition, some construction instances are eliminated as their suitability values have been turned out lower than those of their competitors. As addressed earlier, various factors affect the assessment of suitability. For example, IN_COLOR wins the competition with REL_SVO_WHO and others mostly due to its succinctness – i.e. *woman in blue* is much shorter than *woman who wear blue dress*. Moreover, SVO wins over PAS_SVO since the active voice is more preferable than the passive voice.



(6) This is a snapshot when an "equilibrium state" has reached, where no more meaningful change is detected for construction instances (and obviously in the SemRep). Construction instances that lost in competition have all been eliminated, and the rest of the instances represent the solution that the system came up to for the current SemRep, which is "*woman in blue hit man*".

From a computational point of view, the competition and cooperation paradigm described so far in this section may be understood as a distributed approach applied in the classic search problem. Construction instances can be regarded as independent computing units that encode computational constraints of a certain domain of interaction with a limited scope (i.e. production of linguistic expressions on a part of the SemRep). The invocation process of construction instances roughly corresponds to exploring the search space of the problem, which is in this case, all possible utterances for the given semantics. Similarly, the elimination process corresponds to reducing the search space in a way that the candidates without any possibility leading to the best solution are excluded from the solution pool. The assessed suitability of a construction, or a group of construction, acts as an evaluation function.

However, the amount of the required computation and resources (e.g. the number of construction instances kept in WM, maintenance of connections between instances, the amount of semantic and topographic matching to be done, etc.) in the earlier example may appear to be excessive, especially considering the number constructions invoked for the example (only 16 constructions). In a real situation, when the number of constructions is well above a few hundred thousands, and the scene being described is much more complex, the production process might become intractable due to a combinatorial explosion.

Although there exists the possibility of such an uncontrollable status within the system and it cannot be totally avoided, we would also like to emphasize that it may not be a significant drawback of TCG for a few reasons. Firstly, the process of scene perception is incremental, meaning that the amount of SemRep being built (or updated) at one time is much less than what is shown in the above example (e.g. during simulation described in Section 4.6, the number of constructions updated at once is less than four). Thus, a SemRep of such a size requires several phases to be built, and during those phases, the number of construction instances is likely kept under control (through constant competition processes). Secondly, although the current version of TCG is implemented as a symbolic system, the TCG system can be implemented by using alternative methods, such as a connectionist network. This may alleviate the problems related to huge vocabulary of constructions (e.g. serial access of constructions for matching or invoking). Especially, the storage component for constructions can be implemented in such a way that the access time to the entries is kept constant, independent from the total number of entries stored – e.g. a

Hopfield network (Hopfield, 1982).

4.5. Production Principles of TCG

A task of scene description is highly dynamic and complex in its nature as scene perception and utterance production are interleaving through the process. Especially, the current work addresses the description of natural scenes, and this involves even more complex interactions between perception and production. Even for a static scene – let alone for a dynamic scene, such as a video-clip – a variety of patterns of perception and production can be yielded, depending on diverse situational and behavioral factors. Among other factors, the question of "when" an utterance is being made, or more precisely, how much information is perceived and processed before production of an utterance, seems to be the key factor in deciding the patterns of the outcomes of the process. In fact, very the same question has long been a hotly debated issue in the studies of apprehension and linguistic formulation – it dates back most notably to Lashley (1951) and Osgood (1977) that more recent studies, such as Bock et al. (2004) and Gleitman et al. (2007), discussed with literature reviews and experimental findings (see Section 5.4 for more detailed exposition).

Observations from subjects' utterance production after scene comprehension suggest that even for a similar semantics that the utterances convey, the *well-formedness* of the utterances (e.g. sentence grammaticality, complexity, etc.) varied subject by subject, and case by case. Generally, more well-formed utterances, especially with a complex sentence structure, were produced in less-constrained situations (e.g. under less time pressure) even though there was some degree of subject variability. Although how exactly a well-formed utterance is defined and what kind of properties constitute the well-formedness of an utterance are not yet clear, Section 5.2 and Section 5.3 provide experimental analysis in which a number of different metrics were employed for assessing the well-formedness of utterances from subjects.

This observation leads to the proposal of the *threshold of utterance*, or simply *threshold*, which posits a limit for a speaker to produce an utterance as soon as reached. We propose the threshold of utterance as a theoretical property which sets an upper bound on the "computational resources" spent in interpreting a scene and formulating a sentence before speaking out. This necessitates the distinction of an "utterance" from a "sentence" since the implication of low threshold is that an utterance can be made before completion of a proper sentence. In the current framework of TCG, we define an *utterance* as a group of constructions that are concurrently read out when threshold reaches. Thus, a single utterance may or may not be equated with a single sentence – it may vary from a single word or a phrase to one or more sentences.

The computational resources involved in the production process of TCG generally include the amount of time for building and updating the SemRep and invoking the corresponding construction instances, and the number and the combinatorial depth of activated construction instances within the WM. Simply, threshold in TCG can be reduced to the following formula:

Threshold = min (available Time, available Memory).

The first term (available time) addresses the allowed computational time for producing utterance. In fact, an experiment finding indicates that people generally feel obliged to fill pauses which are getting excessively long, as indicated by their frequent production of verbalized pauses or pause-fillers (e.g. "*uh…*", and "*um…*", etc.) even when they are not under pressure to produce utterances (Section 5.2). The second term (available memory) addresses the upper limit on the available

storage used for formulating utterances, such as working memory or phonological loop. A number of studies reported a fairly limited capacity on both working memory (Lewis, 1996) and phonological loop (Baddeley, 1996, 2003) during linguistic processes. This can be administered by limiting the number of invoked construction instances or the (syllabic) length of the formulating utterance. In combination, both of the terms address the computational overhead and structural complexity that the system is allowed to sustain in formulating the description of a perceived scene.

In real situations, threshold can be set by various factors, such as individual preference (e.g. whether the speaker is talkative or not), scene complexity (e.g. how much time is required to comprehend the scene), task requirements (e.g. whether the speaker is under time pressure to produce utterances), or utterance priority (e.g. an exclamation from the discovery of a surprising fact).

Whenever the system reaches threshold, it is forced to produce an utterance by reading out the mostly "suitable" construction instances, possibly the structure of construction instances with the highest suitability value, at that moment. Given that TCG currently allows an utterance to be produced even before threshold is reached, the produced utterance tends to be more "fragmented" with low threshold (as fewer constructions are available for reading out), whereas high threshold results in more well-formed sentences. Moreover, if the system enters to an "equilibrium state" where there is no more update on construction instances with the SemRep at the moment (i.e. an utterance is "ready"), the system produces an utterance. Thus, the effect of threshold would be more prominent if the sentential structure to be produced is more complex – i.e. for a very simple sentential structure, low threshold would still result in a complete sentence, whereas complex sentences with low threshold may result in ill-formed or incomplete sentence fragments.

One particular thing to note in relation to utterance production in TCG is that the status of an utterance as being "produced" need not be the actual articulation of the utterance. Rather, it is defined as being stored in the phonological output buffer (Jacquemot & Scott, 2006) or as being finished with phonological encoding (Levelt, 1989; Levelt & Meyer, 2000). Thus, the process of reading out addresses the "intention" or the "plan" to produce a certain utterance rather than the state of having finished the production. Although this distinction is not significant in general cases, it may play a crucial role in some cases where specific temporal transitions of the model are important – however, in this thesis, such an aspect is not addressed.

In relation to utterance generation in TCG, there are a few important principles that come into play, especially when various levels of threshold are considered. As briefly mentioned earlier, low threshold may result in a fragmented utterance. In TCG, an utterance can be made "before" the sentential structure is completely prepared or all of its constituents are figured out. This leads to the *premature production* of an utterance (examples in Figure 4.5-1).

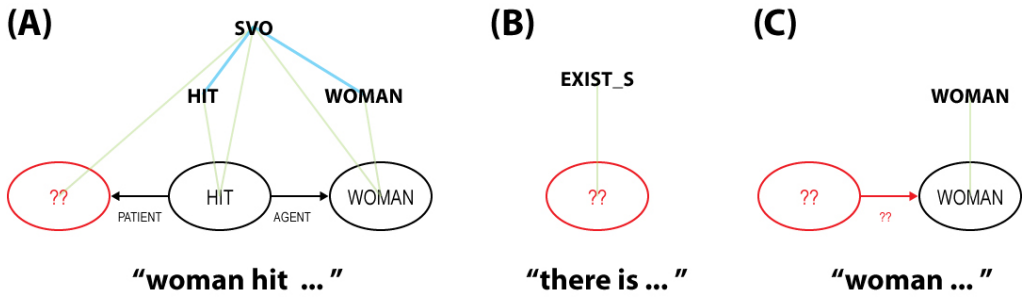


Figure 4.5-1: Example cases of the premature production of an utterance. (A) An utterance is being produced (i.e. constructions are

read out) even though not all constituents are figured out – the patient of the action is still missing (represented as a node in red with an unidentified concept). The system might produce an utterance without fragmentation by inspecting the scene to identify the patient while reading out the prepared portion of the utterance. (B) A similar situation where only the sentential structure is prepared while its only constituent is left unidentified. The system is possibly under the pressure to produce an utterance only after hardly figuring out there is an object, yet still unidentified, in the scene. (C) An opposite case where only the constituent is prepared while the sentential structure is not ready yet.

In fact, a number of studies suggested that speakers do not preplan the whole sentence before the utterance but rather they do it in such an incremental manner that they speak out a partially planned part first and the subsequent planning may continue during the articulation of the first part (Griffin & Bock, 2000; Griffin & Garton, 2003; Griffin & Mouzon, 2004). The range of planning before the sentence onset seems to vary as speakers may preplan for a single phrase structure (R. C. Martin, Crowther, Knight, Tamborello II, & Yang, 2010; R. C. Martin & Freedman, 2001) or until the verb (Bock & Cutting, 1992), or the range may exceed a single syntactic phrase (Schnur, Costa, & Caramazza, 2006) and even span a whole simple sentence (Oppermann, Jescheniak, & Schriefers, 2010). The suggestion is that the scope of preplanning depends on the phonological properties of phrasal words (Schnur, 2011) and their syntactic and thematic relations (Allum & Wheeldon, 2007). Moreover, the prosodic structure of the sentence was also claimed to influence the planning range (Ferreira, 1993; Selkirk, 1984). Griffin (2003) suggested that there is a tradeoff between fluency and incrementality of the production process and speakers try to find the balance point between them – they try to maximize the length of prepared words and the time needed for preparing subsequent words while minimizing the overhead for word buffering. Thus, it seems natural that speakers produce partially prepared utterances, and the premature production principle is supposed to capture such an aspect.

Another important aspect involved in a prematurely produced utterance is that a prematurely produced utterance does not always result in a grammatically or semantically fragmented sentence. Depending on the situation, such an utterance may or may not end up being fragmented, and especially when the prepared portion of the utterance is enough to cover the temporal gap to fill in the missing parts, the utterance may be spoken out smoothly (e.g. A in Figure 4.5-2). The proposed concurrency between the vision and language processes in TCG lays the ground for the smooth development of prematurely spoken utterances as the coordinated processes of vision and language allow the system to articulate an utterance while the vision system gathers information to prepare for the successive utterances.

The implication is, in a more formal statement, that there is the *utterance continuity* in the production process in that the previously produced utterance influences the later process of utterance formulation in a way that those two become a continuous expression as a whole. By this principle, subjects can produce utterances with grammatical continuity even if there are pauses in between (Section 5.2). This is not a special extension from the regular production process of TCG since it can be simply regarded as allowing the construction instances that are already read out to stay a little longer in the WM and to be treated as regular instances in the cooperation and competition process – i.e. they too can be combined with other (newly invoked) instances, forming groups, and participating in competition when there is a conflict. Through this mechanism, for example, the case (A) in Figure 4.5-1 can successfully produce "*woman hit … man*" (if we assume that the newly identified constituent is MAN) rather than the fragmented utterance "*woman hit … there is man*", which is the case where the already

read out constructions, including the sentential structure, are removed from the memory, so a new sentential structure, EXIST_S, is invoked for describing MAN.

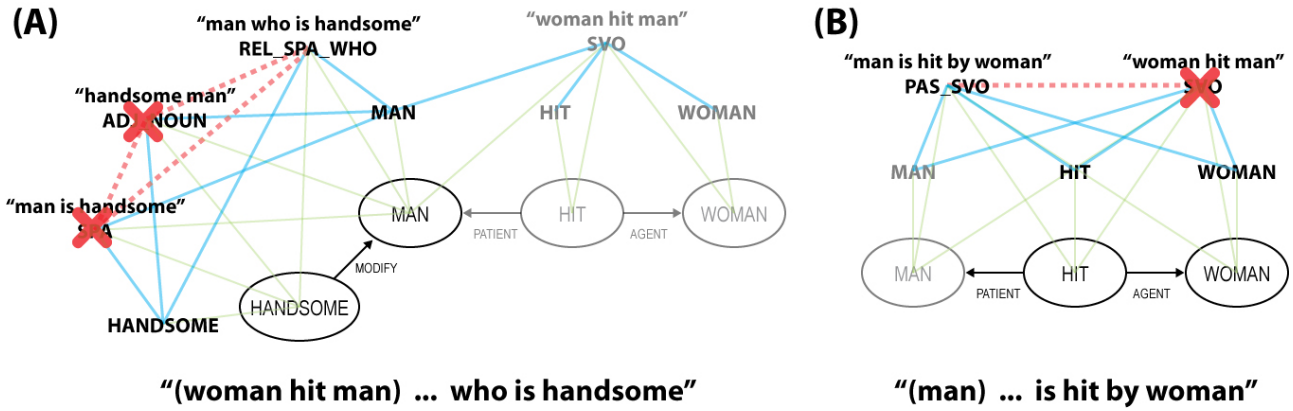


Figure 4.5-2: Example cases of utterances produced in continuum with the previous utterances. (A) REL_SPA_WHO is chosen from the competition because ADJ_NOUN and SPA are not compatible with the syntactic structure of the previously spoken utterance (shown as the shaded part in the SemRep) – i.e. for ADJ_NOUN, *handsome* should have been produced earlier than *man*, and SPA represents another sentence-level structure which cannot be fit in the previous SVO. (B) Only a single construction MAN is uttered, and it biases the sentential structure of the following utterance – without the MAN construction already spoken, SVO (the active voice) would have been chosen over PAS_SVO (the passive voice). PAS_SVO is chosen because it allows the patient of the action, *man*, to be spoken first.

Another principle in the process of TCG addresses the process of the vision system as well – the highly dynamic and interactive nature of scene description process allows the language system to bias the vision system. Such type of bias is classified as the *verbal guidance*, which denotes the case where the utterance structure under formulation guides visual attention. Especially combined with prematurely produced utterances, the constituent being produced next, if unidentified yet, is more likely to be attended firstly among others. The perceptual salience of another constituent may be overridden too (an example in Figure 4.5-3).

In fact, a tight link between visual attention and speech production has been demonstrated from a number of studies: the order of speakers' eye movement directly matched with the order of mentioning of constituents in an adjective noun phrase (e.g. "*the large red ball is next to the mouse*") and a prepositional phrase (e.g. "*the ball, next to the mouse, is large and red*") (van der Meulen, 2001), and speakers looked back at the previously inspected object for describing its property, such as the color, even when the property was not displayed any more (Meyer, Van der Meulen, & Brooks, 2004). Moreover, it has been reported that during the verbal description task, speakers' eye movements generated significant cross-language differences (English speakers tend to attend more on the manner component of the action compared to Greek speakers) even during the first second of motion onset. This was clearly contrast to the free-viewing task, in which speakers allocated attention similarly regardless of the language difference (Papafragou, Hulbert, & Trueswell, 2008). Webb, Knott and MacAskill (2010) also suggested that visual attention may be guided by non-perceptual factors, such as event representations. They reported that during observing a reach-to-grasp action, there was a sequential pattern of saccades, with the agent being fixated first, and then the target, following the order of an active event representation.

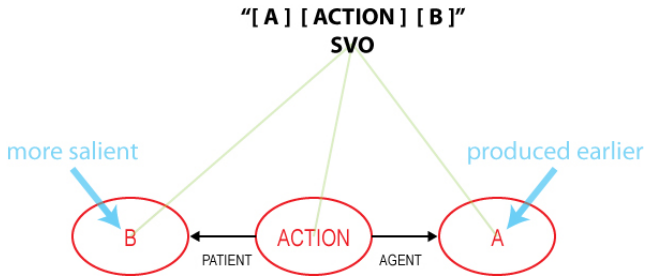


Figure 4.5-3: An example of a verbal representation guiding visual attention. SVO is being produced, or prepared to be produced, prematurely while all the constituents for the agent, the action, and the patient still are not identified enough (shown as red nodes). Since SVO represents an active sentence, the agent, which corresponds to the node A, is more likely to be attended first, even if the patient, which is represented as the node B, is perceptually more salient.

With the above three principles, different levels of threshold together can manifest a variety of utterance and fixation patterns in scene description as highlighted in Chapter 5. Here we provide two example cases where different levels of threshold result in extreme patterns of scene description – the extreme cases of high and low threshold – to demonstrate how the interplay of vision and language works within the current framework of TCG.

In Figure 4.5-4 and Figure 4.5-5, a series of computational stages of TCG for a high and low threshold case are illustrated, which are based on the construction set depicted in Figure 4.3-4 and Figure 4.3-5. Those stages are an abridged version of real simulation results, which highlight important procedural details (see Appendix C for the actual outputs of the implemented system equivalent for each case). For producing these examples, two sets of parameters (i.e. available time and available memory) tuned accordingly for a high and low threshold case, respectively. Moreover, exactly the same SemRep is assumed to be provided as input to the system – i.e. the SemRep is updated with the same elements in exactly the same temporal order and duration for both of the cases. At the top left corner of the figures is the representation of the state of visual attention for generating the SemRep at each stage (a dashed red oval represents a subscene), whose operations are explained in detail in Section 2.7.

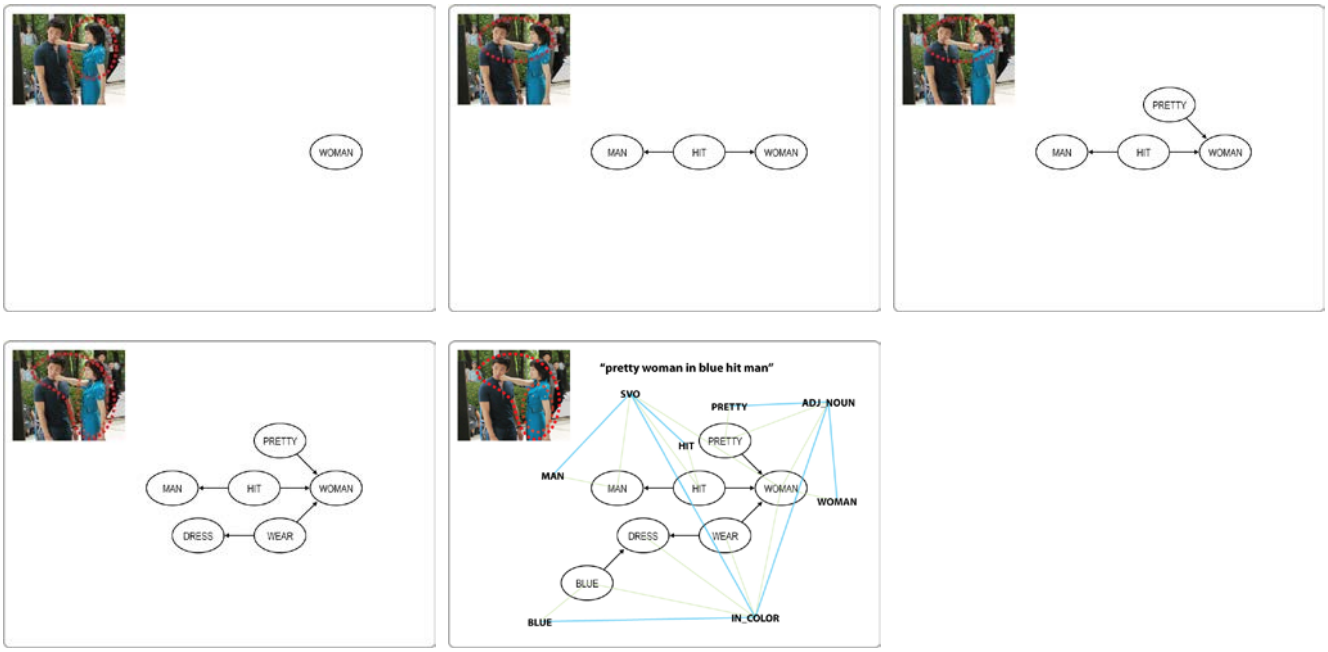


Figure 4.5-4: An example of a high threshold case. High threshold allows the system to have enough time to wait for a big SemRep to be formed and to fully formulate a relatively complex sentential structure accordingly. See text for more detail.

In Figure 4.5-4, a high threshold case is represented as a series of illustrations of each procedural stage. Firstly, the woman is perceived first (in terms of a subscene), and the corresponding SemRep is built. Since threshold is set to be high, the system does not need to produce utterance yet. In the second stage, the initial subscene is extended to include the hitting event happening between the woman and the man (a subscene is built incrementally by "extension" as discussed in Section 2.7). At the third stage, attention zooms into the woman and the detail of the woman, her prettiness, is perceived. Again, the perception is done in terms of a subscene, which is this time as a substructure of the initial subscene (shown as a faded oval). Still no utterance has been made. The fourth stage extends the subscene by perceiving another aspect of the scene, which is the woman's apparel. Next in the final stage, attention zooms in and the color of the woman's dress is perceived and included. Now the utterance "*pretty woman in blue hit man*" is finally produced by reading out constructions applied so far, either due to reaching threshold or achieving an equilibrium state – in the real simulation, the latter reason caused the production since threshold was set well higher than required. Note that throughout the stages of simulation, constructions are constantly applied to the SemRep although the intermediate details (e.g. invocation, cooperation, or competition among construction instances) are omitted from the figures for clarity and only the end result is shown at the last stage.

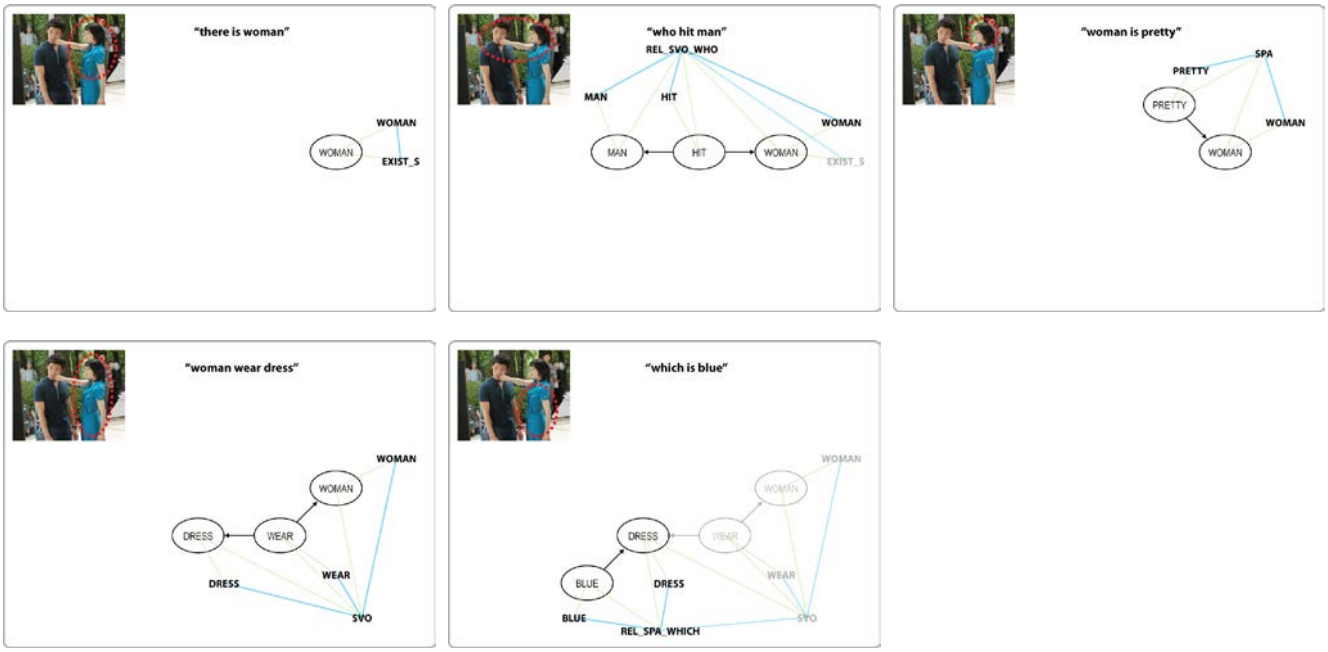


Figure 4.5-5: An example of a low threshold case. Low threshold enforces the system to keep producing utterances, even before the event is fully comprehended. The produced utterances tend to be short and relatively fragmented. See text for more detail.

On the other hand, Figure 4.5-5 shows a low threshold case with a series of illustrations of each procedural stage. Since the SemRep is updated by exactly the same schedule as the high threshold case, at the first stage, the woman is again perceived first (again, in terms of a subscene), and the corresponding SemRep is built. However, in this case, threshold is set to be low, the system is now forced to produce an utterance already, by reading out available constructions at the moment (EXIST_S and WOMAN), which result in the utterance of "*there is woman*". At the second stage, the subscene has been extended to include the hitting event and again the system is forced to produce an utterance. But at this stage, the produced utterance is not an independent sentence but a continuous clause (i.e. "*who hit man*" rather than "*woman hit man*") – the utterance continuity principle. Note that the previously read-out components are shown slightly faded in the figure. At the third and the fourth stage, the SemRep is updated again (the part that is already verbally described is shown faded) and the corresponding utterances are made – in this case, the utterances are independent sentences ("*woman is pretty*" and "*woman wear dress*") since concatenation with the previously spoken utterances is not possible. At the final stage, an utterance is made, again as a continuous clause, and the simulation finishes. Note that in this case of low threshold, a new subscene is perceived at each stage (i.e. no substructure is made inside the previous subscene) as opposed to the high threshold case, where the initial subscene is being extended at every stage while the newly perceived subscenes are included in as substructures (the red ovals with the faded dashed line in Figure 4.5-4).

4.6. Implementation of TCG

This section provides explicit details on the implemented version of TCG, whose theoretical framework is explained in the earlier sections. The implemented model receives a conceptual representation of a perceived scene (through a scene

description file) as input, which describes regions of events and objects with the associated perceptual schemas assumed to be perceived from the scene, and generates utterances that describe the semantics of the perceived scene. The system simulates the production process of TCG as described in Section 4.4 while applying the production principles introduced in Section 4.5. A primitive version of the vision system is also implemented in the model in order to simulate the attention mechanism for perception of the scene (through regions defined in the scene description file). The user can set a number of parameters, such as threshold, or simulation time.

A. Development

The model is implemented in C++ by using Microsoft Visual Studio 2008 and the simulation results were produced on Windows 7 platform. However, the code for the implemented model is written following the convention of the standard C++ (e.g. GCC), and the application is designed to run on the command prompt. Thus, the implemented model can be compiled and run on other platforms, such as UNIX, or MacOS.

B. Architecture

The following provides specific details of the architecture and the relevant components of the implemented model of TCG. Although most of them sincerely follow the theoretical framework of TCG discussed so far, some differ in certain details and some have extra features that are not addressed.

Visuo-Linguistic Working Memory

The most basic component of the implemented model of TCG is the *Visuo-Linguistic Working Memory (VLWM)* (see Figure 4.3-1 for the schematic view of the system). The VLWM provides a workspace where: the SemRep is formulated, construction instances are invoked, construction instances cooperate and compete with each other, and the verbal description of a perceived scene is generated.

The VLWM includes the term of vision and language within its name because it contains all types of element necessary for the task of scene description. The VLWM contains nodes and relations of the current SemRep (in the implementation, they are defined as *SemRep instances*), construction instances, and combined structures of construction instances that represent partially or fully complete utterance fragments (i.e. construction structures). SemRep instances and construction instances are assigned with activation levels. They remain (or they are maintained) in the VLWM until the activation levels drop to 0. Note that different from SemRep and construction instances, construction structures are stored in the VLWM temporarily as they are all reset at each simulation time – they are used only temporarily for the competition and the utterance production process.

Semantic Network

The *semantic network* of the implemented model of TCG provides the basic definition of all the semantic concepts and categorical knowledge required in running the system – it defines the semantic meanings of all of the concepts specified in perceptual schemas perceived from a scene, SemRep elements built from perceptual schemas, constructions defined in the repertoire, and construction instances resulted from the invocation process.

In theory, such a network is supposed to be implemented as a type of the schema network of conceptual knowledge.

However, the current implementation of the semantic network is reduced to a set of literal symbols that represent conceptual meanings and categories (e.g. ENTITY, HUMAN, WOMAN, MAN, etc.) and a set of relations between those symbols (e.g. TURTLE is-a ANIMAL). Similar to a typical semantic network, concepts in the current implementation of the semantic network are defined in terms of relations with other concepts. Currently, only "*is-a*" relation is used, but more relation types will be included as more examples with complex constructions are covered by the model.

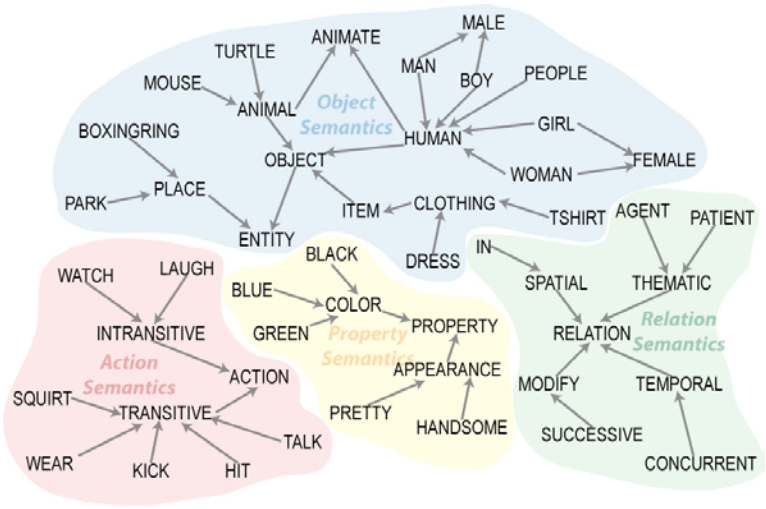


Figure 4.6-1: A schematic view of the semantic network used during the simulations presented in the current work. Only "is-a" relation (depicted as an arrow) between conceptual entries is used.

The semantic network is stored in the system in a form of graph structure where each node represents a concept and an edge, which is directed (from the subordinate level concept to the superordinate level concept), represents a relation between concepts. Due to its simplicity, the semantic process in the semantic network is reduced to the simple logical inference between conceptual entries (i.e. concepts represented as literal symbols). Categorical judgment between concepts is simply done by checking a relational connection between them – i.e. check if traversing from a concept to another concept is possible. For example, according to Figure 4.6-1, WOMAN is in the category of HUMAN (connection exists) but not in the category of MALE (no connection exists). In the current implementation, similarity between concepts, which can be done by measuring the relational distance between concepts, is not seriously considered during matching two concepts (e.g. invocation process) because only the concepts with exactly the same symbol (i.e. the relational distance of 0) are supposed to be "matched". However, if either of the comparing concepts is defined as "inclusive" (as denoted by an attached "+" sign), all of the subordinate category level concepts are judged to be matched – e.g. according to Figure 4.6-1, OBJECT+ is matched all with HUMAN, MAN, and DRESS.

Scene Perception

Since TCG is supposed to address the dynamics of scene perception and the corresponding utterance production, the utterance production processes of TCG described in the earlier sections assumes a tight correlation with mechanisms of the vision system (e.g. the verbal guidance principle). Thus, it is necessary to include a certain type of vision process that can

establish the inter-related functional link between the vision and language systems, and the current implementation of TCG contains “preliminary” mechanisms of the vision system.

The implemented mechanisms of the vision system mostly concern attention shifts during scene perception. The system receives a conceptual description of a perceived scene, which consists of the regions of the scene and the associated perceptual schemas. Currently, a region is defined as an oval of a certain size (i.e. vertical and horizontal radii) at a certain location (specified in terms of the center point).

Except for the description of the associated perceptual schemas and the location and size information, each region has two parameters: *saliency* and *uncertainty*. The saliency of the region specifies how perceptually salient the region is, whereas the uncertainty specifies how “difficult” to perceive the region. The system places attention on one of the regions which has the highest saliency value unless another region is biased to be attended first (e.g. the verbal guidance principle drives attention to the region associated with the construction instance to be produced next). The “inhibition of return” is applied so that already perceived region is not going to be attended again. The uncertainty is defined in terms of the required simulation time until which the inspection on the region has to be maintained to perceive the region – e.g. a region with the uncertainty of 2 requires attention to stay on that region for 2 simulation times to be perceived. Thus, a region with uncertainty of 0 is perceived immediately without the necessity of placing attention on that region, implying that the region is considered to provide the gist of the scene. When a region is “perceived”, the associated perceptual schemas are regarded to be perceived as well. The system creates or updates SemRep elements according to the type of perceived perceptual schemas.

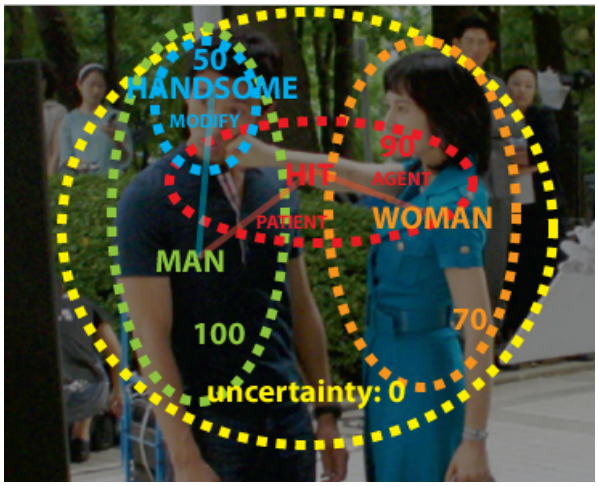


Figure 4.6-2: A schematic view of an example scene description where 5 regions are defined. Some regions are associated with perceptual schemas – object schemas are represented as big capital letters (e.g. HANDSOME, WOMAN, HIT) and relation schemas are shown as smaller capital letters over faded lines connecting objects schemas (e.g. PATIENT, MODIFY). Numerical values represent the saliency of regions. The yellow region, whose uncertainty is 0, represents the gist of the scene (the uncertainty of other regions is all set to 1).

Perceptual schemas associated with a region specify the semantics that the region represents – e.g. a region covering the face of a man in a scene may be associated with perceptual schemas that specify the properties of the man’s face, such as his handsomeness. We currently define two types of perceptual schemas for the implemented system: *object schema* and *relation*

schema. An object schema is defined by the associated region and the concept that it conveys. A relation schema too is defined by the associated region and the concept, but it also contains links to the object schemas that it connects (i.e. specifies the relation between those schemas). Note that the concepts specified within perceptual schemas are the ones defined in the semantic network used by the system.

Object and relation schemas are directly mapped to the nodes and relations of SemRep, respectively. When system perceives an object schema, a SemRep node is created in the VLWM of the system whereas the perception of a relation schema yields a new relation. If the node or relation associated with the perceived schema already exists, only update on that component is made. Therefore, by deploying attention to the regions specified in the scene description provided to the system, perception of the scene – i.e. building the SemRep – is performed. The scene description provides dynamics in SemRep formulation, in which the saliency of the region specifies the order in creating (and updating) SemRep components in the VLWM while the uncertainty defines the delay during the process.

Invocation of Construction Instance

When there is a new or an updated piece of the SemRep in the VLWM, a number of construction instances are "invoked". All of the constructions defined in the vocabulary of the system are matched with the updated part of the SemRep. When a construction is judged to be "matched" – i.e. the topology of the Sem-Frame of the construction overlaps and the concepts of the corresponding SemRep elements are judged to be matched after consulting the semantic network (see Section 4.4 for the detailed conditions) – a construction instance is invoked over that area of the SemRep. The invoked construction instance "covers" the region of the SemRep by being associated with the components (i.e. nodes and relations) of the covering area of the SemRep. Currently, the invocation process is being done only over the updated region of the SemRep – a construction instance can be invoked over a wider region than the updated area, but the covering area has to contain at least one updated or newly created element.

Once invoked, a construction instance is assigned with an activation level and remains in the VLWM of the system until the activation level drops to 0. Construction instances with the activation levels of 0 are eliminated from the VLWM.

Competition and Cooperation Paradigm

The theoretical description of TCG provided in the earlier sections does not address the specific details on how the competition and cooperation paradigm should be implemented. There are broadly two types of implementation style that are possible as shown in Figure 4.6-3.

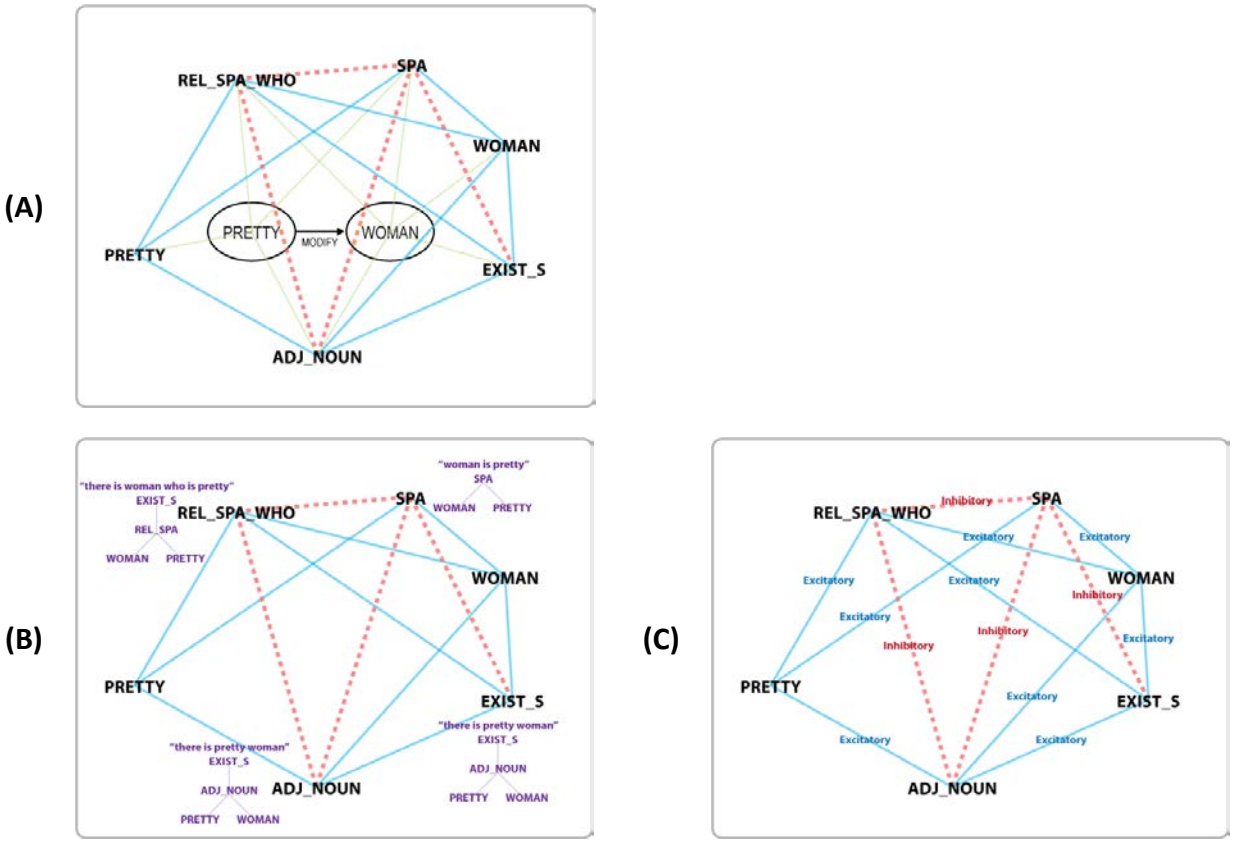


Figure 4.6-3: Illustrations of two possible approaches for implementing the competition and cooperation processes between construction instances. (A) shows construction instances invoked over a SemRep. These instances cooperate (shown as the blue lines) and compete (shown as the red lines) with each other. (B) illustrates an algorithmic approach where all possible cooperative combinations between construction instances (i.e. construction structures) are considered and suitability of each instance is assessed based on those combinations. The combination with the maximum suitability is selected during competition between construction instances (e.g. for the REL_SPA construction, the construction structure for "*there is woman who is pretty*" is selected as the most suitable one). (C) illustrates another approach which is based on the network of cooperative and competitive connections between construction instances. The solution of the system (i.e. produced utterance) is achieved by the convergence through a number of iterative computations.

One approach, which has been taken for the current version of implementation, is based on an algorithmic method that searches all the possible connections between cooperating construction instances. During the cooperation process, the system builds all possible *construction structures* that are resulted from combinations between construction instances. A construction structure is a set of construction instances that meet the requirements for the combination through slots, such as class matching, overlapping in covering regions (see Section 4.4 for the detailed conditions). The created construction structures are (temporarily) stored within the VLWM of the system with other construction instances and SemRep elements. Each of the construction structures represents a partial syntactic structure of the utterance being formulated (i.e. an utterance fragment), which is considered to be one possibility to the final solution of the system (i.e. produced utterance). Among the instances SPA, PRETTY, and WOMAN in Figure 4.6-3, for example, a total of 6 combinations are possible (i.e. SPA, WOMAN,

PRETTY, SPA-PRETTY, SPA-WOMAN, SPA-WOMAN-PRETTY) and the system creates 6 construction structures (although the first five would be removed immediately since they are subparts of the last one). Each of the created constructions structures is assessed for its suitability, and the structure with the highest suitability is chosen as the solution when threshold reaches (i.e. read out). When construction instances compete, each of the instances is assigned with the maximum suitability among all of the construction structures that it belongs to (i.e. the best solution is selected), and the instance with lower suitability decreases its activation level and eventually eliminated – currently, the construction instance that loses competition is immediately eliminated without setting its activation level to 0. In this approach, the competition and cooperation between construction instances are done in terms of construction structures as each of them represents the "global suitability" of an instance assuming that it ends up in that particular structure.

On the other hand, another approach, which calculates the suitability of each construction instance based on the cooperative and competitive network, is also possible. In this approach, construction instances form a network by having cooperative (excitatory) and competitive (inhibitory) connections with other instances when the requirements for the cooperation and competition are met (see Section 4.4 for the detailed conditions). The strength of each connection made between construction instances is proportional (whether it is excitatory or inhibitory) to the suitability of each construction instance. Through these connections, the activation level of each construction instance converges as excitatory connections gradually increase the activation level while inhibitory connections decrease it. The system iterates for a reasonable number of computation phases (e.g. 10,000 times of iteration) and the activation level of each construction instance keeps adjusting according to the number and type of connections that it has during the iteration. Construction instances with activation levels below a certain level (say, 0) would be eliminated and the remaining instances are chosen for the solution. In this approach, the competition and cooperation between construction instances are done simultaneously through the network connections.

In the current implementation, the first approach has been taken because the range of the scene description task that the current framework of TCG addresses requires considerations on the global structure of the produced utterance – e.g. the utterance continuity principle requires the sentential structure of the to-be-produced construction instances since the construction structure of a grammatically continuous structure with the previously produced utterance is assigned with higher suitability than others with uncontinuous structures. Taking the global features into account in the second approach is extremely difficult because such features are an emergent outcome of distributed computation through the network. Moreover, the second approach is practically difficult to manage as convergence is not always guaranteed. The whole process is very sensitive to the setting of the parameter values, such as the initial activation levels of construction instances, or connection weights, and sometimes the system does not reaches to a stable state even after a fairly large number of iterations (e.g. oscillation happens).

Assessment of Suitability

When construction instances are combined to yield a construction structure, the suitability of that structure is calculated. The suitability of a construction structure represents how "suitable" the construction instances within the structure as a whole are to be chosen as the produced utterance. After the cooperation process, each of the construction instances within the VLWM of the system is assigned with the maximum suitability selected from all of the construction structures that the

instance is registered as a member. Suitability is used for competition between construction instances (i.e. the instance with higher suitability wins) and selection of the construction structure to be uttered (i.e. the structure with the highest suitability is chosen).

The suitability of a construction instance is calculated by the following formula:

Suitability = (# of non-shared covering SemRep elements) × W_S – (# of syllables in phonetic notations) × W_L + (preference value) × W_P.

W_S, W_L and W_P all represent weight values that are set to 100, 1, and 50, respectively. As clearly seen from the formula, suitability is proportional to the covering area of the SemRep (i.e. wider covering area means more semantics is represented) and the preference value, whereas it is inverse-proportional to the length of verbal expression to take the ease of production into account (i.e. longer expression is harder to articulate). The suitability of a construction structure is simply a sum of the suitability of all of its member instances.

When the utterance continuity principle (Section 4.5) is set to be applied, the suitability of each of the construction structures is adjusted in the way that the structures that grammatically conform to the previously produced utterance are "rewarded" while the structures with un-continuous syntactic structures are "penalized". In the current implementation, if the beginning of an utterance is exactly overlapped with the ending of another utterance and the syntactic structure of one of those utterance is totally included in that of the other, those two utterances are considered to be *grammatically continuous* – e.g. the utterance "*woman hit man*" and "*man who is handsome*" are grammatically continuous. The number of overlapped "syntactic components" between the previously read out construction structure and the current construction structure (i.e. the overlapped construction instances and the cooperative connections between those instances) is counted, and it is used to adjust the suitability by the following formula:

Adjustment (for continuous structures) = Suitability + (# of overlapped syntactic components) × W_R.

Adjustment (for uncontinous structures) = Suitability - (# of overlapped syntactic components) × W_R.

W_R represents a weight value for redundancy, which is set to 100. Thus, a construction structure (and a construction instance) is penalized or rewarded as much grammatical redundancy as it has compared to the already produced utterance. For example, if the previously produced utterance was "*woman hit man*", the utterance "*woman is pretty*" will be more preferable to the utterance "*pretty woman hit man*" since both of them are not grammatically continuous but the former (only *woman* is redundant) has less redundancy than the latter (*woman*, *hit*, and *man* are redundant).

Utterance Production

When threshold is reached, the system produces an utterance by reading out the construction structure with the highest suitability. Since a construction structure is a type of tree structure, the system produces an utterance by visiting the construction instances in inorder. When an empty slot is reached during the traversal (i.e. the slot of the visiting construction instance is not connected to another instance), the system stops utterance production at the moment. If some utterance has been made, an ellipsis ("…") is appended to indicate that there is more utterance to be produced left. If no utterance has been produced before the process halts due to an empty slot, the system produces a pause-filler ("*uh*…") rather than remaining silent. When the verbal guidance principle is in effect, the scene region associated with the missing constituent is assigned as

the region to be attended in the next simulation time. This is possible since a region is associated with perceptual schemas which spawn SemRep elements when perceived, and these SemRep elements are associated with construction instances when being covered while some of them are linked with slots in the Syn-Form of the constructions.

Once an utterance is produced, the construction instances in the read-out construction structure and the elements of the SemRep that those instances are covering are all marked as *old*. Old instances and SemRep elements are soon to lose their activation levels and will be eliminated unless read out again. Currently, old instances and elements drop their activation levels to 0 at the next simulation time, and the instances and elements with the activation levels of 0 are eliminated from the VLWM during the maintenance process. Thus, the instances and elements marked as old can survive at least for "one" simulation time. Producing an utterance resets the activation levels of all of the construction instances in the read-out construction structure and the SemRep elements that they cover, lengthening their life time for one simulation time. Note that the reason of keeping old construction instances (and their covering SemRep elements) in the VLWM at least for one simulation time is to support the utterance continuity principle.

Threshold of Utterance

In Section 4.5, we have provided a formal definition of the threshold of utterance and a simple formula defined in terms of available computational resources. As indicated by the formula, threshold is bound by two aspects of computational resources, time and memory, and we define the following three parameters for threshold in the current implementation.

- The simulation time elapsed since the last production of utterance.
- The total number of syllables that can be kept in the system – i.e. the sum length of the syllables of all of the construction instances in the VLWM. Note that a slot is counted as 0 syllable.
- The total number of construction instances in the VLWM.

The first one addresses the time aspect and the latter two address the memory aspect. Whenever the system reaches the upper limit set by any of the three parameters, the system is forced to produce an utterance. If there is no available construction structure at the moment, the system skips utterance production.

The system also produces utterance even if threshold is not reached when there is no more update on the SemRep (i.e. no more scene region is perceived).

Simulation Cycle

The current implementation of TCG uses a "relative" time frame in which a single cycle of processes completes one simulation time. Since it is relative, one simulation time is not directly matched with any real time unit, such as a second, or a minute, but rather it generally corresponds to a cognitively important transition of the conceptual status of a speaker. Thus, when compared to the real data, the order of events is preserved, but the relative temporal duration is not necessarily matched – e.g. an event which takes two simulation times does not mean that it is twice as long in real time as another event taking a single simulation time. In fact, the relative length of each event (e.g. perception of certain information from a scene) is totally dependent on how long the user defines the event to take, especially by assigning the uncertainty value for a scene region.

In the current implementation, a single simulation cycle performs as follows:

1) The system performs the maintenance process, during which construction instances and SemRep elements (i.e. nodes and relations of the SemRep) whose activation levels are 0 or below are eliminated. Construction instances that lost competition during the previous simulation time are eliminated too. “Old” construction instances and SemRep elements also drop their activation levels to 0, but it happens “after” the elimination phase so that they stay in the VLWM until the next simulation cycle.
2) The vision process is done by deploying attention and perceiving a scene. Attention shifts to the region with the highest saliency unless forced to move to a certain region (e.g. by the verbal guidance principle). Attending to a region lowers its uncertainty, and when it becomes 0 perception of the region is performed, which results in updating the SemRep.
3) Construction instances are invoked over the updated areas of the SemRep.
4) The cooperation process is performed, during which construction instances combine with each other to create construction structures (all possible combinations are considered), and their suitability is assessed. Before the cooperation process begins, all of the previously made construction instances are removed from the VLWM.
5) Construction instances compete with each other when there is conflict between them. The loser construction instances (i.e. the ones with lower suitability than their competitors) are marked to be eliminated. Constructions structures that contain the loser construction instances are eliminated too.
6) The system produces an utterance when threshold is reached by reading out the construction structure with the highest suitability among all the construction structures created at the current simulation time.
7) The internal state of the system (construction instances, SemRep elements, competition trace, etc.) is printed out.
8) The next simulation cycle begins – go back to (1).

The simulation stops if there is no more activity in the VLWM and no more regions to perceive. It also stops when the simulation time exceeds a preset value.

Production Principles

The production principles explained in Section 4.5 are also implemented in the system. Each of them can be set on and off before simulation begins.

The premature production principle is implemented in such a way that before utterance production, the system checks the construction structure selected to be read out to see if it contains a missing slot. If the premature production principle is set to “off” and not all of the slots of the selected construction structure are filled in, the production of utterance at that simulation time is skipped.

The utterance continuity principle is implemented simply by allowing the adjustment of suitability to be effective or not. If the utterance continuity principle is set to “on”, the suitability of construction structures is adjusted according to their grammatical continuity with the previously produced utterance.

The verbal guidance principle is simply implemented by directing attention to the region that is associated with the missing slot of the read out construction structure. At the next simulation time, attention is placed on that region, allowing the system to update the SemRep to eventually invoke construction instances that will fill in the slot.

C. Data Format

The current implementation model receives three types of input. They are provided in a text format. The system reads the data files for the semantic network, construction vocabulary, and the scene description file, and produces the simulation output, which is also in a text format.

Before we take a look at the data files and their formats used in the system, there are several things to keep in mind. Firstly, when reading data files, the system does not distinguish line-change characters from other blank letters, thus equating a multi-line statement to a single-line statement. Moreover, all blank letters (e.g. spaces, tabs, etc.) are treated as the same, and the number of blanks does not count. This means that all of the following statements are treated as the same statement.

`is_a ANIMATE { HUMAN ANIMAL }`	`is_a ANIMATE`
	`{`
	`  HUMAN`
	`  ANIMAL`
	`}`
		`is_a`
		`ANIMATE {`
		
		`HUMAN ANIMAL }`

Secondly, data files can contain comment lines, and a comment line is denoted as the symbol “#”. Any letters that come after the symbol are skipped until a new line starts.

The following is an excerpt from the data file that defines the semantic network (Figure 4.6-1) used in the system (see Appendices for the complete data used for the simulations presented in the current work) – the excerpted part defines the object-related semantics.

```
# object-related semantics
is_a ENTITY
{
        OBJECT
        {
                HUMAN
                {
                        MAN BOY
                        WOMAN GIRL
                        PEOPLE
                }
                ITEM
                {
                        CLOTHING
                        {
                                DRESS
                                TSHIRT
                        }
                }
                ANIMAL
                {
                        MOUSE TURTLE
                }
        }
        PLACE
```

```
        {
                BOXINGRING
                PARK
        }
}
is_a ANIMATE { HUMAN ANIMAL }
is_a MALE { MAN BOY }
is_a FEMALE { WOMAN GIRL }
```

As specified earlier, only “is-a” relation is implemented. The *is_a* keyword shown above specifies the “is-a” relationship between a superordinate concept and its subordinate concepts. The convention of the statement is as follows:

is_a [superordinate concept] { [subordinate concept 1] [subordinate concept 2] … }.

Thus, the statement “is_a HUMAN { MAN WOMAN }” defines two “is-a” relations, MAN is-a HUMAN and WOMAN is-a HUMAN. The statement can be written recursively as follows:

is_a [concept A1] { [subordinate concept B1 of concept A1] { [subordinate concept C1 of concept B1] … } … }.

Note that the keyword *is_a* is used only at the beginning of the statement and it is not necessary again within the recursive field. Thus, the statement “is_a OBJECT { HUMAN { MAN WOMAN } }” defines three relations, HUMAN is-a OBJECT, MAN is-a HUMAN and WOMAN is-a HUMAN.

Another input data file required in the system is the construction vocabulary file, which defines all constructions used for running simulation. The following is an excerpt from the vocabulary file (see Appendices for the complete data used for the simulations presented in the current work), which defines the IN_COLOR construction.

```
# the definition of IN_COLOR construction
construction IN_COLOR
{
    class: NP

    node HUMAN { concept: HUMAN+ shared head }
    node WEAR { concept: WEAR }
    node CLOTH { concept: CLOTHING+ }
    node COLOR { concept: COLOR+ shared }
    relation HUMAN_WEAR { concept: AGENT from: WEAR to: HUMAN }
    relation CLOTH_WEAR { concept: PATIENT from: WEAR to: CLOTH }
    relation COLOR_CLOTH { concept: MODIFY from: COLOR to: CLOTH }

    [HUMAN: NP N] 'in' [COLOR: A]
}
```

The name of the construction (in this case, IN_COLOR) is given next to the keyword *construction* in the first line. The keyword *node* and *relation* are used to define a node and a relation that belong to the Sem-Frame of the construction, and the name of a node or a relation comes after the corresponding keyword. The names are necessary because they are used as pointers in the definition of a node, which needs the names of nodes that it connects, and a slot, which is linked to a Sem-Frame element. A node or a relation also contains a concept that it represents. The keywords *shared* and *head* specify whether the element is defined as “shared” or the head of the construction.

The definition of a slot specifies the name of a Sem-Frame element that it is linked with and a list of the classes of other constructions that can fill in the slot. The convention of the definition is as follows:

[(Sem-Frame element name) : (class 1) (class 2) (class 3) …].

A phonetic notation is defined between apostrophes (“”). The order between slots and phonetic notations are preserved in the Syn-Form of the construction – what is defined first is considered to be what is going to be produced first during utterance production.

The last input data file provided to the system is the scene description file, which defines the regions and perceptual schemas that are going to be perceived during simulation. The following is an excerpt from the scene description file that defines two regions (the yellow and red region) depicted in Figure 4.6-2.

```
# the gist of the scene
region GIST
{
        location: 216, 117 size: 150, 100
        saliency: 0             # saliency doesn't matter
        uncertainty: 0          # instantly perceived

        # layout
        perceive WOMAN = ENTITY, HIT = ACTION, MAN = ENTITY
        perceive HIT_AGENT, HIT_PATIENT
}

# hitting area
region HIT_AREA
{
        location: 213, 110 size: 60, 20
        saliency: 90
        uncertainty: 1

        object HIT { concept: HIT }
        relation HIT_AGENT { concept: AGENT from: HIT to: WOMAN }
        relation HIT_PATIENT { concept: PATIENT from: HIT to: MAN }

        perceive HIT, HIT_AGENT, HIT_PATIENT
}
```

The name of a region is defined next to the keyword *region* (in this case, GIST and HIT_AREA). Each region has fields for location, size, saliency, and uncertainty, which are specified by the corresponding keywords. Except for these fields, the definition of a region can also contain definition of the perceptual schemas associated with the region, whose definition convention is very similar to the definition of the Sem-Frame of a construction described above. Note that a perceptual schema (whether it is an object or relation schema) can be referred to from other regions (e.g. HIT, HIT_AGENT, and HIT_PATIENT schemas, which are defined in the region HIT_AREA, are referred to in the region GIST).

The keyword *perceive* specifies the perceptual schemas that are going to be “perceived” when attending to the region is finishes during simulation. The concept of the referred perceptual schema is allowed to be replaced by a temporary concept in

order to support the gist perception mechanism. During perception of the gist of a scene, some objects in the scene might not be fully identified, and such ambiguity in the identification is implemented as replacing the concept of a perceived perceptual schema with a superordinate level concept – e.g. in the region GIST, the concept of WOMAN and MAN schema are temporarily replaced with ENTITY, implying that WOMAN and MAN are not fully identified at the moment when the GIST region is perceived. The convention of the definition of perceived schemas is as follows:

perceive [perceptual schema 1], [perceptual schema 2], [perceptual schema 2] …

or,

perceive [perceptual schema 1] = [replacing concept for perceptual schema 1], [perceptual schema 2] = [replacing concept for perceptual schema 2] …

Upon receiving the types of data files given above, the system runs simulation and produces output. Currently, the output consists of the iterations of simulation time, in which four types of information are represented: current attention location, VLWM status (the current SemRep and construction instances), competition history, construction structures, produced utterance, and next attention location. The following is an example output of the system (only simulation time 3 is shown).

```
================================================================================
  Simulation Time: 3
================================================================================
> Current Attention
  KICK_AREA (perception done)

> Perceived Regions
  KICK_AREA

> Schema Instances
[ @] SemRep-N BOY_0
[ @] Construction EXIST_S_1 covering BOY_0 for 'there is' [REL_PAS_SVO_WHO_11]
[ @] Construction BOY_2 covering BOY_0 for 'boy'
[!@] SemRep-N KICK_3
[!@] SemRep-R PATIENT_4 from KICK_3 to BOY_0
[!@] SemRep-R AGENT_5 from KICK_3 to HUMAN_6
[!O] SemRep-N HUMAN_6
[!X] Construction SVO_7 covering HUMAN_6 BOY_0 KICK_3 AGENT_5 PATIENT_4 for [ ] [KICK_12] [BOY_2]
[!X] Construction PAS_SVO_8 covering HUMAN_6 BOY_0 KICK_3 AGENT_5 PATIENT_4 for [BOY_2] 'is' [KICK_12]
'-ed by' [ ]
[!O] Construction EXIST_S_9 covering HUMAN_6 for 'there is' [REL_SVO_WHO_10]
[!X] Construction REL_SVO_WHO_10 covering HUMAN_6 BOY_0 KICK_3 AGENT_5 PATIENT_4 for [ ] 'who' [KICK_12]
[BOY_2]
[!@] Construction REL_PAS_SVO_WHO_11 covering HUMAN_6 BOY_0 KICK_3 AGENT_5 PATIENT_4 for [BOY_2] 'who
is' [KICK_12] '-ed by' [ ]
[!@] Construction KICK_12 covering KICK_3 for 'kick'

> Competition Traces
  SVO_7(343) eliminated PAS_SVO_8(287)
  SVO_7(343) eliminated REL_SVO_WHO_10(333)
  REL_PAS_SVO_WHO_11(627) eliminated SVO_7(343)
```

```
> Construction Structures
[X] 343: SVO_7 [ ] [KICK_12 'kick'] [BOY_2 'boy']
[X] 287: PAS_SVO_8 [BOY_2 'boy'] 'is' [KICK_12 'kick'] '-ed by' [ ]
[*] 627: EXIST_S_1 'there is' [REL_PAS_SVO_WHO_11 [BOY_2 'boy'] 'who is' [KICK_12 'kick'] '-ed by' [ ]]
[X] 333: EXIST_S_9 'there is' [REL_SVO_WHO_10 [ ] 'who' [KICK_12 'kick'] [BOY_2 'boy']]

> Produced Utterance
  "who is kick-ed by..."

> Next Attention
  GIRL_AREA (uncertainty left: 1)
```

The *Schema Instances* field shows the content of the VLWM at the current simulation time – with construction instances, the elements of the SemRep are also treated as schema instances as *SemRep-N* represents a node of the SemRep and *SemRep-R* represents a relation of the SemRep. A schema instance is represented by its name attached with an ID number (e.g. KICK_3) whose name is set to be the meaning of the concept for a SemRep element and the name of the construction for a construction instance. In front of each instance, the status of an instance is shown, which is represented by a combination of symbols that are specified as follows:

- "!" – the instance is updated or newly created.
- "O" – the instance is in the normal condition.
- "X" – the instance is dead (soon to be eliminated).
- "@" – the instance is old.

The *Competition Traces* field shows the record of all competitions (between construction instances) that happen during the current simulation time. The numeric value appears next to the name of a construction instance is the suitability that the instance is assigned (a construction instance can belong to multiple construction structures and the maximum suitability among them is chosen).

The *Construction Structures* field lists the construction structures that are left after the pruning process has been done. Construction structures are initially created based on all possible combinations among construction instances at the moment, but most of them are immediately "pruned" if they are a subpart of other construction structures. Only a few are left after the pruning process, and each of those left construction structures represents a unique syntactic structure (e.g. a different root or combination of construction instances). In front of each construction structure is the status of the structure, which is represented by a combination of symbols that are specified as follows:

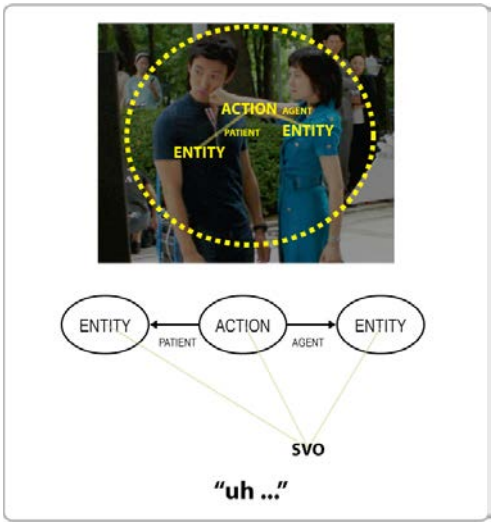
- "X" – the structure contains a dead construction instance (i.e. invalid construction structure).
- "*" – the structure is selected to be read out for producing an utterance (i.e. the highest suitability).

The numeric value appears next to the status is the suitability of the construction structure, and what appears next is the content of the structure, the recursively represented connectivity of the construction instances in the structure.

D. Simulation

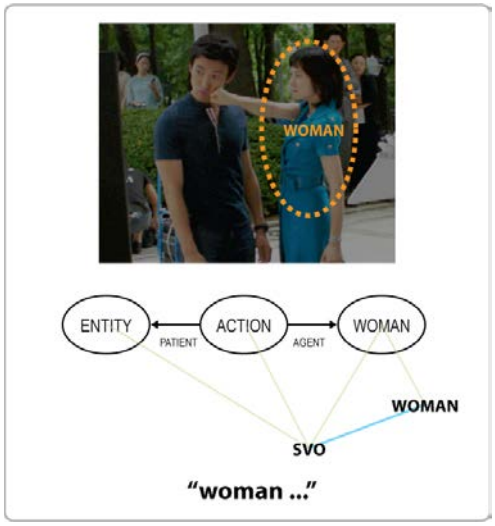
Now we turn our focus to the specific steps, through which simulation is performed in the current implementation of TCG. The following is a series of illustrations of an example simulation, which is a type of low threshold case (the time

parameter is set to 1, and the number of construction instances and syllables are set to infinite). All of the production principles are set to be effective, and the scene description file corresponding to Figure 4.6-2 is provided to that system (see Appendices for the entire simulation result).



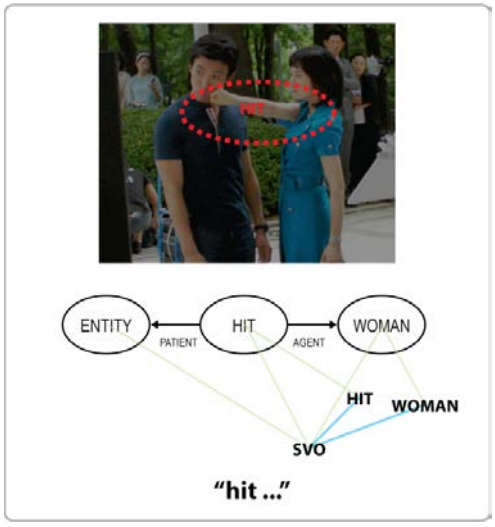

```
> Current Attention
  None

> Perceived Regions
  GIST

> Schema Instances
[!O] SemRep-N ENTITY_0
[!O] SemRep-N ACTION_1
[!O] SemRep-N ENTITY_2
[!@] SemRep-R AGENT_3 from ACTION_1 to ENTITY_0
[!@] SemRep-R PATIENT_4 from ACTION_1 to ENTITY_2
[!@]  Construction  SVO_5  covering  ENTITY_0  ENTITY_2  ACTION_1
AGENT_3 PATIENT_4 for [ ] [ ] [ ]
[!X] Construction PAS_SVO_6 covering ENTITY_0 ENTITY_2 ACTION_1
AGENT_3 PATIENT_4 for [ ] 'is' [ ] '-ed by' [ ]

> Competition Traces
  SVO_5(250) eliminated PAS_SVO_6(194)

> Construction Structures
[*] 250: SVO_5 [ ] [ ] [ ]
[X] 194: PAS_SVO_6 [ ] 'is' [ ] '-ed by' [ ]

> Produced Utterance
  "uh..."

> Next Attention
  WOMAN_AREA (uncertainty left: 1)
```

Simulation Time 1: The layout of the event is provided first, and based on that the SVO construction is initially invoked. Since the system has to produce utterance in every single simulation time (the threshold time is set to 1), the system tries to produce an utternace. Currently, only "*uh…*" is produced because the first slot of SVO is empty – the premature production principle allows such an incomplete utterance to be produced. Due to the verbal guidance principle, the next attending region is set to WOMAN_AREA, which is associated with the missing first slot of SVO. According to the scene description provided to the system (although not shown currently), the region with the highest saliency is MAN_AREA.



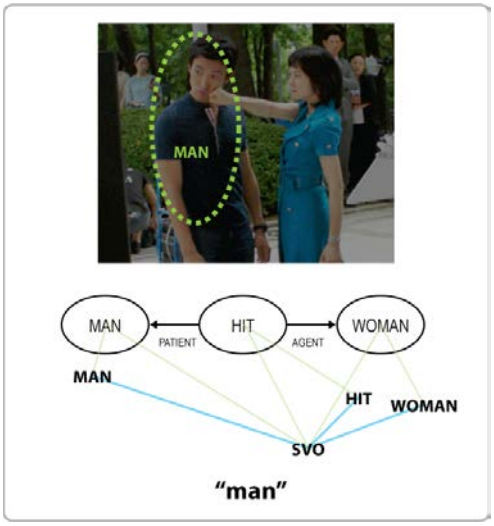

```
> Current Attention
  WOMAN_AREA (perception done)

> Perceived Regions
  WOMAN_AREA

> Schema Instances
[!@] SemRep-N WOMAN_0
[ O] SemRep-N ACTION_1
[ O] SemRep-N ENTITY_2
[ @] SemRep-R AGENT_3 from ACTION_1 to WOMAN_0
[ @] SemRep-R PATIENT_4 from ACTION_1 to ENTITY_2
[ @] Construction SVO_5 covering WOMAN_0 ENTITY_2 ACTION_1
AGENT_3 PATIENT_4 for [WOMAN_11] [ ] [ ]
[!X] Construction PAS_SVO_8 covering WOMAN_0 ENTITY_2 ACTION_1
AGENT_3 PATIENT_4 for [ ] 'is' [ ] '-ed by' [WOMAN_11]
[!O] Construction EXIST_S_9 covering WOMAN_0 for 'there is'
[REL_SVO_WHO_10]
[!X] Construction REL_SVO_WHO_10 covering WOMAN_0 ENTITY_2
ACTION_1 AGENT_3 PATIENT_4 for [WOMAN_11] 'who' [ ] [ ]
[!@] Construction WOMAN_11 covering WOMAN_0 for 'woman'

> Competition Traces
  SVO_5(445) eliminated PAS_SVO_8(289)
  SVO_5(445) eliminated REL_SVO_WHO_10(335)

> Construction Structures
[*] 445: SVO_5 [WOMAN_11 'woman'] [ ] [ ]
[X] 289: PAS_SVO_8 [ ] 'is' [ ] '-ed by' [WOMAN_11 'woman']
[ ] 138: EXIST_S_9 'there is' [WOMAN_11 'woman']
[X] 335: EXIST_S_9 'there is' [REL_SVO_WHO_10 [WOMAN_11 'woman']
'who' [ ] [ ]]

> Produced Utterance
  "woman..."

> Next Attention
    HIT_AREA (uncertainty left: 1)
```

Simulation Time 2: The system produces the utterance "woman" as WOMAN_AREA and the associated WOMAN schema are perceived, updating the WOMAN node in the SemRep. Again, the next region to be attended is set to HIT_AREA, not MAN_AREA, due to the verbal guidance principle.



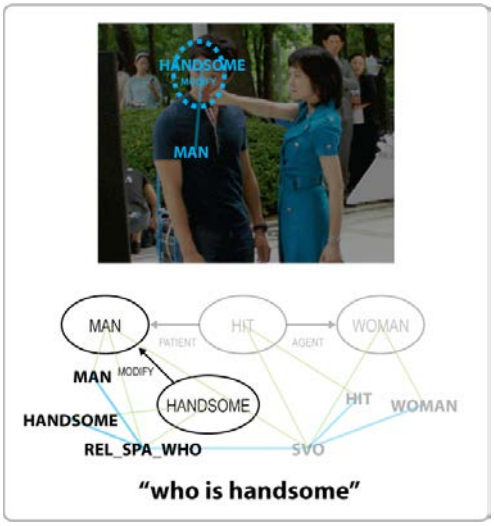

```
> Current Attention
  HIT_AREA (perception done)

> Perceived Regions
  HIT_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[!@] SemRep-N HIT_1
[ O] SemRep-N ENTITY_2
[!@] SemRep-R AGENT_3 from HIT_1 to WOMAN_0
[!@] SemRep-R PATIENT_4 from HIT_1 to ENTITY_2
[ @] Construction SVO_5 covering WOMAN_0 ENTITY_2 HIT_1 AGENT_3
PATIENT_4 for [WOMAN_11] [HIT_15] [ ]
[ O] Construction EXIST_S_9 covering WOMAN_0 for 'there is'
[REL_SVO_WHO_14]
[ @] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[!X] Construction PAS_SVO_13 covering WOMAN_0 ENTITY_2 HIT_1
AGENT_3 PATIENT_4 for [ ] 'is' [HIT_15] '-ed by' [WOMAN_11]
[!X] Construction REL_SVO_WHO_14 covering WOMAN_0 ENTITY_2 HIT_1
AGENT_3 PATIENT_4 for [WOMAN_11] 'who' [HIT_15] [ ]
[!@] Construction HIT_15 covering HIT_1 for 'hit'

> Competition Traces
  SVO_5(742) eliminated PAS_SVO_13(286)
  SVO_5(742) eliminated REL_SVO_WHO_14(332)

> Construction Structures
[*] 742: SVO_5 [WOMAN_11 'woman'] [HIT_15 'hit'] [ ]
[X] 286: PAS_SVO_13 [ ] 'is' [HIT_15 'hit'] '-ed by' [WOMAN_11
'woman']
[X] 332: EXIST_S_9 'there is' [REL_SVO_WHO_14 [WOMAN_11 'woman']
'who' [HIT_15 'hit'] [ ]]

> Produced Utterance
  "hit..."

> Next Attention
  MAN_AREA (uncertainty left: 1)
```

Simulation Time 3: HIT_AREA and the associated HIT schema are perceived, updating the HIT node and the two relations in the SemRep, eventually invoking the HIT construction instance that can fill in the second slot of SVO. The system now produces the corresponding utterance, "hit …"




```
> Current Attention
  MAN_AREA (perception done)

> Perceived Regions
  MAN_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ @] SemRep-N HIT_1
[!@] SemRep-N MAN_2
[ @] SemRep-R AGENT_3 from HIT_1 to WOMAN_0
[ @] SemRep-R PATIENT_4 from HIT_1 to MAN_2
[ @] Construction SVO_5 covering WOMAN_0 MAN_2 HIT_1 AGENT_3
PATIENT_4 for [WOMAN_11] [HIT_15] [MAN_21]
[ O] Construction EXIST_S_9 covering WOMAN_0 for 'there is'
[REL_SVO_WHO_19]
[ @] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ @] Construction HIT_15 covering HIT_1 for 'hit'
[!X] Construction PAS_SVO_17 covering WOMAN_0 MAN_2 HIT_1
AGENT_3 PATIENT_4 for [MAN_21] 'is' [HIT_15] '-ed by' [WOMAN_11]
[!O] Construction EXIST_S_18 covering MAN_2 for 'there is'
[REL_PAS_SVO_WHO_20]
[!X] Construction REL_SVO_WHO_19 covering WOMAN_0 MAN_2 HIT_1
AGENT_3 PATIENT_4 for [WOMAN_11] 'who' [HIT_15] [MAN_21]
[!X] Construction REL_PAS_SVO_WHO_20 covering WOMAN_0 MAN_2
HIT_1 AGENT_3 PATIENT_4 for [MAN_21] 'who is' [HIT_15] '-ed by'
[WOMAN_11]
[!@] Construction MAN_21 covering MAN_2 for 'man'

> Competition Traces
  SVO_5(1039) eliminated PAS_SVO_17(283)
  SVO_5(1039) eliminated REL_SVO_WHO_19(329)
  SVO_5(1039) eliminated REL_PAS_SVO_WHO_20(323)

> Construction Structures
[ ] 140: EXIST_S_18 'there is' [MAN_21 'man']
[*] 1039: SVO_5 [WOMAN_11 'woman'] [HIT_15 'hit'] [MAN_21 'man']
[X] 283: PAS_SVO_17 [MAN_21 'man'] 'is' [HIT_15 'hit'] '-ed by'
[WOMAN_11 'woman']
[X] 329: EXIST_S_9 'there is' [REL_SVO_WHO_19 [WOMAN_11 'woman']
'who' [HIT_15 'hit'] [MAN_21 'man']]
[X] 323: EXIST_S_18 'there is' [REL_PAS_SVO_WHO_20 [MAN_21
'man'] 'who is' [HIT_15 'hit'] '-ed by' [WOMAN_11 'woman']]

> Produced Utterance
  "man"

> Next Attention
  MAN_FACE_AREA (uncertainty left: 1)
```

Simulation Time 4: The system completes the full SVO sentence by producing the utterance “man” after perceiving MAN_AREA.




```
> Current Attention
  MAN_FACE_AREA (perception done)

> Perceived Regions
  MAN_FACE_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ @] SemRep-N HIT_1
[ @] SemRep-N MAN_2
[ @] SemRep-R AGENT_3 from HIT_1 to WOMAN_0
[ @] SemRep-R PATIENT_4 from HIT_1 to MAN_2
[ @] Construction SVO_5 covering WOMAN_0 MAN_2 HIT_1 AGENT_3
PATIENT_4 for [WOMAN_11] [HIT_15] [REL_SPA_WHO_25]
[ O] Construction EXIST_S_9 covering WOMAN_0 for 'there is' [ ]
[ @] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ @] Construction HIT_15 covering HIT_1 for 'hit'
[ O] Construction EXIST_S_18 covering MAN_2 for 'there is'
[ADJ_NOUN_26]
[ @] Construction MAN_21 covering MAN_2 for 'man'
[!@] SemRep-N HANDSOME_22
[!@] SemRep-R MODIFY_23 from HANDSOME_22 to MAN_2
[!X] Construction SPA_24 covering MAN_2 HANDSOME_22 MODIFY_23
for [MAN_21] 'is' [HANDSOME_27]
[!@] Construction REL_SPA_WHO_25 covering MAN_2 HANDSOME_22
MODIFY_23 for [MAN_21] 'who is' [HANDSOME_27]
[!X] Construction ADJ_NOUN_26 covering MAN_2 HANDSOME_22
MODIFY_23 for [HANDSOME_27] [MAN_21]
[!@] Construction HANDSOME_27 covering HANDSOME_22 for
'handsome'

> Competition Traces
  REL_SPA_WHO_25(1326) eliminated SPA_24(237)
  SPA_24(237) eliminated ADJ_NOUN_26(232)

> Construction Structures
[X] 237: SPA_24 [MAN_21 'man'] 'is' [HANDSOME_27 'handsome']
[ ] 227: EXIST_S_18 'there is' [REL_SPA_WHO_25 [MAN_21 'man']
'who is' [HANDSOME_27 'handsome']]
[X] 232: EXIST_S_18 'there is' [ADJ_NOUN_26 [HANDSOME_27
'handsome'] [MAN_21 'man']]
[*] 1326: SVO_5 [WOMAN_11 'woman'] [HIT_15 'hit']
[REL_SPA_WHO_25 [MAN_21 'man'] 'who is' [HANDSOME_27
'handsome']]
[X] 131: SVO_5 [WOMAN_11 'woman'] [HIT_15 'hit'] [ADJ_NOUN_26
[HANDSOME_27 'handsome'] [MAN_21 'man']]

> Produced Utterance
  "who is handsome"

> Next Attention
  None
```

Simulation Time 5: As the details of the man's face (the HANDSOME node and the modifying relation) has been specified by perceiving MAN_FACE_AREA, the system now produces the clause "*who is handsome*" in continuation of the already produced sentence "*woman hit man*". In this case, the utterance continuity principle plays a significant role as other possible syntactic structures, such as "*man is handsome* (the first construction structure)" and "*there is handsome man* (the third construction structure)", are all eliminated. They score significantly low suitability (237 and 232) compared to the selected construction structure (1326) since their grammatical continuity has been considered.

Although it is still in a preliminary stage, the example simulation illustrated so far demonstrates how much of variety the production principles of TCG and the different levels of threshold together can manifest in the patterns of fixation and utterance. Chapter 5 addresses this issue in more detail with findings from eye-tracking experiments. The key idea of the study is that a different level of threshold may result in an utterance with a different degree of well-formedness – high threshold tends to produce relatively well-formed utterances while low threshold is more likely to produce more fragmented utterances. In fact, evidence suggests that threshold induced by time pressure may affect the well-formedness of speakers' utterance (Section 5.3), as demonstrated by the following utterances collected from the actual speech made by subjects during an eye-tracking experiment.

Subject JI, Low Threshold Case

```
um
there are two women
one of them is wearing a really big
dress that's green
and she is kicking the other woman
who is wearing a blue dress
and
sh-
this looks like some kind of boxing match
because they're in a ring
and there are people watching them
```

Subject KF, High Threshold Case

```
a woman in a green dress
is kicking
a woman in a blue dress
in
what looks like a boxing ring
with many people watching the show
```

The grammatical competence, or well-formedness, of the above utterances is quite different – JI's utterances are relatively fragmented, which are mainly short sentences and clauses interleaved with pauses, whereas KF's utterances are comparatively "intact" as they form a single complex sentence with a few embedded clauses. The difference in the experimental settings applied on each case is time pressure elicited by task requirements – the former case required subjects to produce utterances as quickly as possible while the latter did not impose any speed requirements.

Although the above addresses only a particular example and the current implementation of TCG is yet far from mimicking the performance of human speakers, it should be worthwhile to run simulation that results in the contrasting utterance pattern highlighted in the above example.

In the following (Figure 4.6-4 and Figure 4.6-5), we provided visualized illustrations of two simulation results (see Appendices for the actual simulation outputs) where high and low threshold induced by different levels of time pressure result in utterances of different degrees of well-formedness, similar to what is shown above. Thus, only the time parameters are tuned accordingly – for thigh threshold, the time parameter is set to infinite, whereas for low threshold, it is set to 1. The other parameters, the number of construction instances and syllables, are set to infinite. All of the production principles (utterance continuity, premature production and verbal guidance) are set to be effective. The same scene description, which is based on the scene used in the eye-tracking experiment that the subjects JI and KF participated in, is provided to both of the cases.

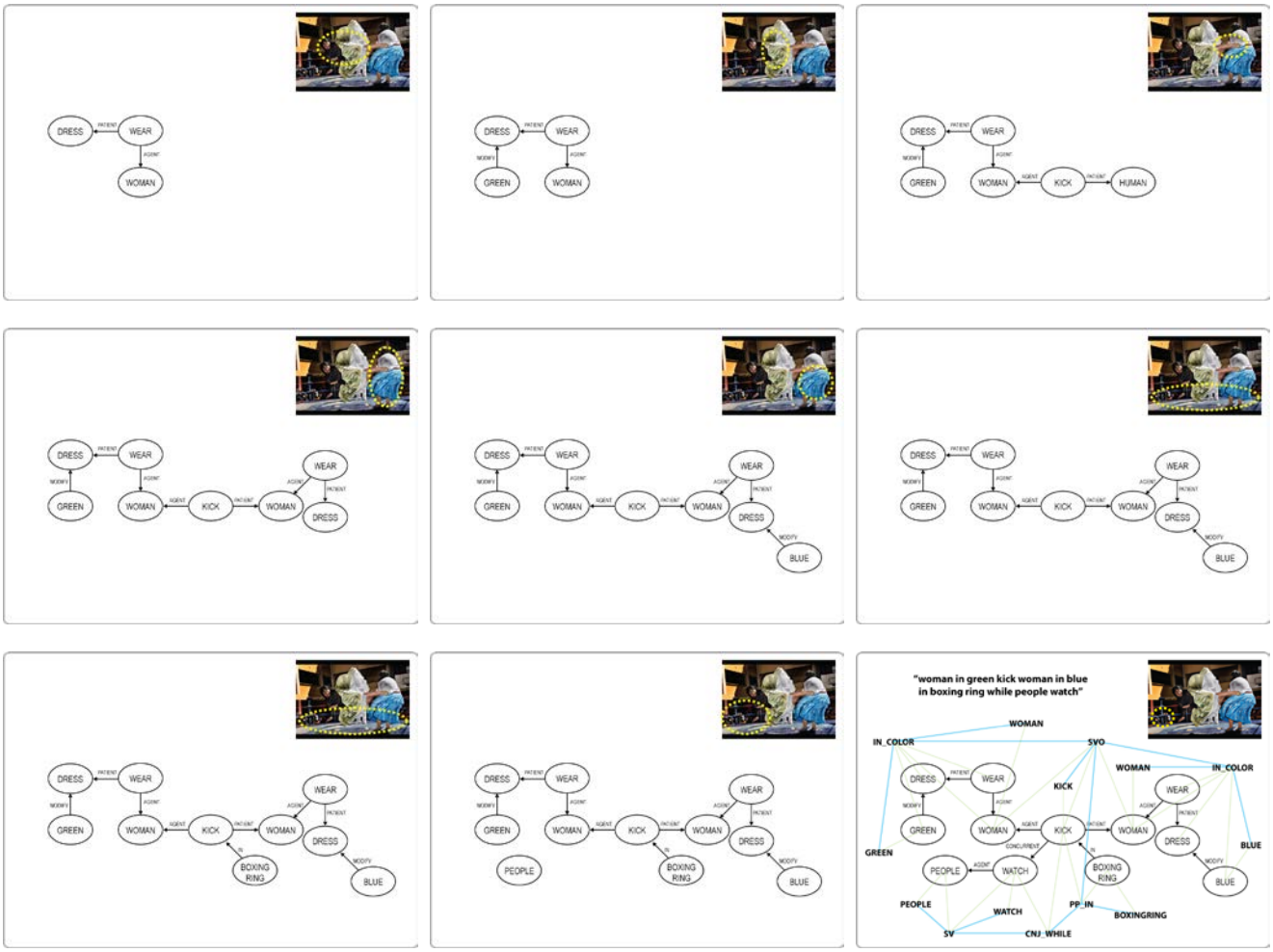


Figure 4.6-4: An illustration of the simulation result of a high threshold case. High threshold allows the system to have enough time to wait for a big SemRep to be formed and to fully formulate a relatively complex sentential structure accordingly. The yellow oval with a dashed line represents the location of attention at the moment (as specified by the attended region of the scene). Note that the uncertainty of almost all of the regions is set to 1, resulting in attention shifting to another region at every simulation time (except for the boxing ring area attended at simulation time 7 and 8, which is set to 2).

Figure 4.6-5: An illustration of the simulation result of a low threshold case. Low threshold enforces the system to keep producing utterances, even before the event is fully comprehended. The produced utterances tend to be short and relatively fragmented, especially due to the elimination of the SemRep that has been described – e.g. at simulation time 7 and 9, newly created nodes and relations are not properly connected with the previous SemRep.

Simulated Utterance, Low Threshold Case

```
woman wear dress
which is green
woman kick...
woman who wear dress
which is blue
it is boxing ring
there is people
who watch
```

Simulated Utterance, High Threshold Case

```
woman in green kick woman in blue in boxing
ring while people watch
```

As shown above, the produced utterances from the simulation for both of the cases are similar to the actual utterances from subjects – the simulated utterances with low threshold are relatively fragmented short clauses and sentences (but connected smoothly in a grammatically appropriate manner as the utterance continuity principle is in effect), whereas the

utterance with high threshold is a well-formed single sentence with a few embedded clauses. As we intended in designing the scene description provided to the system, the order and the semantics of the simulated utterances well match with the real utterances.

However, the scene description, which specifies the dynamics of eye movements and the availability of scene semantics, is designed in an ad-hoc manner – the fixation patterns of speakers' eye movements gathered from eye-tracking experiments are not thoroughly considered. In future work, general patterns of subjects' fixations need to be analyzed, and the simulation results should provide a wide variety of cases with different conditions to explain the experiment data.

4.7. Other Language Models

We are currently implementing a parsing for our own version of construction grammar, Template Construction Grammar (TCG). In some sense, TCG may be seen as a variant of other computational models of Construction Grammar. However, TCG exhibits a number of distinctive characteristics. Firstly, TCG grounds its approach to language by using SemRep as the format for its semantics, which is explicitly designed to link the semantics of sentences to the representation of visual scenes. The use of SemRep involves a sufficiently general graphical structure that we are confident of its extensibility to other meanings. Moreover, each concept in a SemRep, which is itself an abstraction of schema assemblages, is associated with a perceptual schema whose processing is claimed to be instantiated in neural activities, which lays groundwork for multi-modal integration across sensory and motor systems (Section 2.1). Especially, cooperative computation through direct simulation of schemas (i.e. the competition and cooperation processes among schema instances) is proposed to capture neural activities in the brain at a functional and structural level (Michael A. Arbib, 1981; Michael A. Arbib, et al., 1998). Furthermore, our emphasis on the Construction Grammar framework lies in the aspect of constructions acting as the abstract structural template of syntax and semantics, which eventually allows TCG to handle complex sentential structures of multi-level hierarchy, such as sentences with embedded clauses (Section 4.3).

One effort to implement a computational model based on the Construction Grammar formalism is Fluid Construction Grammar (FCG) (De Beule & Steels, 2005; Steels & De Beule, 2006a, 2006b). According to Steels (2006b), FCG is a fully operational formalism for Construction Grammar, which utilizes a uniform mechanism for parsing and production, implying that all the rules and constructions are "bi-directionally" defined, hence usable both for parsing and production. As it is expected from the name, "Fluid" Construction Grammar, the process is proposed to be highly flexible in the sense that it can cope with partially ungrammatical or incomplete sentences. The focus of FCG is on the origins of linguistic evolution based on the premise that language users constantly change and update their grammars. FCG adopts predicate structure for representing constructions and their (syntactic and semantic) constraints while using logical deduction as the basis for processes of comprehension and production.

In contrast to TCG, however, there are a few shortcomings in the approach of FCG. Firstly, due to the nature of its approach, FCG uses complex representations (i.e. logical predicates) with multiple structure types, which renders FCG very difficult and unintuitive to use – constructions in FCG are basically defined in a form of various types of rules, each of which is defined as an exact transformation process between the verbal expression and the meaning, in order to capture the categorical divergence in the semantic and syntactic hierarchy. On the other hand, TCG adopts a single format for all types of

constructions (although we conceptually separate constructions into simple and complex) regardless of the syntactic or semantic level of a construction. And the representational format of the semantic and syntactic structure of a construction (i.e. the Sem-Frame and the Syn-Form), as well as their application rules, are much simpler and more intuitive. Another shortcoming of FCG from the Cognitive Grammar point of view is that FCG is inherently symbolic, leaving the formalism not best suitable for addressing how embodied knowledge is related to process of language understanding.

Another approach which seeks to place Construction Grammar in a computational framework related to an agent's interaction with the external world is Embodied Construction Grammar (ECG) (Bergen & Chang, 2005). ECG is a comprehension model which adopts the basic constructionist definition of a grammatical construction, but emphasizes the relation of constructional semantic content to embodiment and sensorimotor experiences. A central claim is that the content of all linguistic signs involve mental simulations and are ultimately dependent on motor-schema-like entities, which are called X-schemas (Narayanan, 1997, 1999). Although ECG tries for an embodied approach in language understanding, 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.

In contrast to TCG, again, there are a few shortcomings in the approach of ECG. ECG is simpler than FCG in its format of constructions, but it also taps on different construction 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 the categorical hierarchy in the semantics and syntax of language, and this requires multiple constructions and schemas defined in advance in order to define a new construction entry. In TCG, the whole process of inheritance, if there is any, is reduced to the definition of the concept and class, resulting in a much simpler definitional and representational format. Moreover, ECG does not address the issue of the sentential hierarchy explicitly as the given example only covers a simple ditransitive event – "*Mary tossed me a drink*".

Moreover, there are a few language models of Construction Grammar based on a connectionist approach. Among others, Dominey and colleagues (2006; 2009) proposed a production model based on the role of corticostriatal function in sentence comprehension and non-linguistic sequencing. Especially, this model relies on a neural network constrained by the cortico-striato-thalamo-cortical (CSTC) neuroanatomy of the human language system, which connects Brodmann Area (BA) 47, caudate, thalamus and BA 45 and BA44/6. This CSTC circuit is implemented as a recurrent neural network (RNN), which learns construction patterns based on structural cues – i.e. closed class words. The model inserts lexical semantic information (represented as open class words) into the learned sentential structures to produce sentences. This model specifies the roles of various neural areas in relation with the components of the model in very fine detail, but it fails in providing an account on how sentence with grammatical hierarchy can be processed – the examples addressed by the model are simple transitive sentences.

Similarly, Chang and colleagues (2006) proposed a model which makes use of error-based learning to acquire and adapt sequencing mechanisms and meaning-form mappings to derive syntactic representations. This model is also implemented as an RNN, whose operation, which is to learn the syntactic structure, is supported by a separate route of neural network that

processes the semantic aspect of the language. The XYZ structure, which may be regarded as a type of a construction, has been proposed for mapping between the event semantics and the sentential structure. Although this model can address a number of psycholinguistic findings, especially in related to the priming effect, it also fails in providing an account on grammatical hierarchy.

Except for the shortcomings in dealing with complex sentences with embedded clauses, the connectionist approaches taken by the above models have another problem – they assume that the system is already given with the syntactic categorical structure of the language. Dominey et al.'s model receives input words during the learning phase through two separate routes, one for function words and the other for content words, but it fails to provide specific accounts on how the system is capable of distinguishing function words (or morphemes) from content words. Similarly, Chang et al.'s approach also assumes such a separation in the input to the model, as they provide the event semantics (in a form of the XYZ structure) directly to the model. This is a necessary condition because what their RNNs learn is the structures of syntactic (and/or sematic) categories of various sentence types rather than the sequences of actual words. As it has been claimed that RNNs can only learn the sentence structure with already learned words (if there is a new word in the input sentence, then they fail to learn the structure) (van der Velde, van der Voort van der Kleij, & Kamps, 2004), it should have been inevitable for them to provide the categorical information to their models. However, the crucial piece is still missing in their approaches since the mechanism for extracting the categorical information from words is not explicitly specified (see Miikkulainen & Dyer, 1991 for a previous effort to develop "invariant" categorical representations based on a connectionist network).

Lastly, the U-Space model (Vosse & Kempen, 2000, 2009) appears to be worth mentioning here as TCG and the U-Space model share a number of common properties that are interesting to compare although the U-Space model is a comprehension model. Despite of the different linguistic frameworks of TCG and the U-Space model (U-Space model is based on a generative grammar framework, the head-driven approach), they both operate on the competition and cooperation paradigm – they both focus on the activation, or frequency-based competition, between alternative attachment possibilities offered by syntactic building blocks retrievable from the mental lexicon. These blocks are defined as construction instances in TCG, whereas they are defined as lexical frames in the U-Space model. Lexical frames are partial-tree-like structures with syntactic configurations only. They can be connected to other lexical frames during comprehension process, eventually forming a full parse tree of the input sentence – this is what the model produces as an interpretation of the sentence. Similar to TCG, lexical frames cooperate (by forming links between lexical frames) and compete (by inhibiting links of other conflicting lexical frames) with each other. During the comprehension process, links are made between lexical frames if their syntactic structures are compatible, and only a few links are left at the end of the process as other links are eliminated due to the inhibition process between those links. The model produces a parse tree by simply following those links, and that is considered to be the interpretation of the current sentence.

However, TCG and U-Space model differ in a few important ways. Firstly, the grammatical approach adopted by the U-Space model (Performance Grammar) is "lexicalized" in such a way that the information needed to build grammatically correct sentences is claimed to be associated with the individual lexical items. This is in contrast with a Construction Grammar approach taken in TCG, which is more "global" in the way that a single construction may cover a sentential structure, or even a bigger structure in a discourse level. This lexicalized approach limits the U-Space model's capability of

parsing to be “myopic”, especially when dealing with patterns that stretch over a number of lexical items, such as idioms. For example, the expression like *kick the bucket* needs to be interpreted as a whole since a regular verb argument, where the *bucket* is treated as the direct object of the verb *kick*, cannot properly interpret the meaning. The U-Space model adopts an ad-hoc method to address such expressions – it solves the problem by tuning parameters in advance for specific links between lexical frames in each case. The downside of this method is that those parameters need to be set again every time the model deals with a different expression. This leads the parsing process of the U-Space model to be very sensitive to different types of parameters, such as the initial link strengths, activation rate, or decay constant. A subtle change in those parameters results in a very different outcome. Thus, the model runs after setting up the parameter values (by simulated annealing) for specific sentence types that the model will parse – if the sentences are changed, then the model needs to set up those parameters again. This makes it impossible for the model to build a stable repertoire of lexical items (i.e. vocabulary) that can be applied to general cases since the model needs to be set up with parameters at every single time when a new type of sentence is to be parsed.

Another property of the U-Space model, which differs from TCG, is that no semantics is considered during parsing sentences. At the end of parsing, the judgment on the correctness of interpretation is done only by inspecting the resultant parse tree. It is one of the most notable properties of the generative grammar framework, but semantics plays an important role even for parsing a simple sentence. Sentences with global ambiguity, such as “*the woman hits the man with stick*”, may be such an example. The prepositional phrase *with stick* can be interpreted as either modifying *the man* (low attachment) or describing the instrument of the action *hits* (high attachment). Without semantic information, the judgment cannot be made – in this case, the high attachment case seems semantically more plausible. The U-Space model solved this problem, again, by manipulating parameter values on the link between *with* and *hits*, but this is still ad-hoc because the model cannot address the opposite case (the low attachment case) without adjusting the parameter values and running again. Although it is beyond the coverage of the current work, TCG may solve such a problem by putting the global-level semantics in consideration since the processes of TCG are intrinsically based on semantics (i.e. SemRep). For example, we can just define two different prepositional constructions defined specific for a hitting action with appropriate semantic variations in the meaning of the second object (i.e. the object comes after the word *with*) – for high attachment, “HittingAction Object *with* InstrumentalObject”, and for low attachment, “HittingAction Object *with* NoninstrumentalObject”. The semantic concept of “InstrumentalObject” and “NoninstrumentalObject” differ in the way that the former is defined as a type of an object that is generally used as a tool of a stretched length (e.g. stick, bat, racket, hammer, etc.) while the latter is an object that lacks such a property. This type of approach is possible in TCG since constructions are defined in terms of both the semantic and syntactic structure.

Chapter 5. Interplay Between Eye Movements and Speech

5.1. Interplay of Vision and Language

Given its dynamic nature, analyzing the task of scene description requires a detailed assessment of the interplay between the vision and language systems. A valuable window on this interplay is provided by the relationship between eye fixations and the related utterances. We have investigated the time course of this relationship to gain insight into the nature of the internal representation being formed in the speaker's mind as well as the cognitive processes that the speaker undergoes. In this chapter, we describe two eye-tracking experiments that we designed and conducted in an effort to test our hypotheses on how semantic representation is built from acquired visual information and how it influences the produced utterances.

Before describing these experiments, we first review a number of key aspects of TCG and SemRep that we asserted when introducing a computational model of scene description in the earlier chapters. These aspects were proposed to capture the subtlety of the interactive processes linking vision and language.

One of the highlighted aspects regarding the perception of a scene and the successive formation of a SemRep is the notion of subscene (Section 2.7). A subscene is defined as a cognitive construct that captures a partial view of the scene covered by a cognitively significant event and entities. A subscene may come in various event types and covering areas, representing a particular interpretation of the scene at a certain moment. An important point is that a subscene is perceived and encapsulated into a SemRep through different procedural steps depending on the perceptual and conceptual properties of the scene.

More specifically, we proposed two scenarios in scene perception as illustrated in Figure 2.7-3, in which the coverage of the immediately perceived subscene plays a crucial role in the diversity of the process. The initial coverage of a perceived subscene depends on whether or not a certain event (or gist) of the scene is immediately recognizable – if an event is easily recognizable, then the event layout (e.g. hitting event) is immediately perceived and is encapsulated into a SemRep but with some details missing, whereas if an event is not immediately recognizable, a subscene is formed on a smaller and more easily recognizable region (e.g. man's face), resulting in a more detailed SemRep but with a smaller covering area. The event of the scene may be more fully perceived later as successive fixations fill in the missing details of the already created event structure, or the event structure is incrementally figured out as more constituents are discovered by successive fixations. Thus, depending on the immediate availability of the layout of the event, subscenes covering the same event are perceived in broadly two styles: the case of subscene "specification" where a subscene is perceived by filling in details, and the case of subscene "extension" where a subscene is perceived by extending its extent in an incremental manner.

Therefore, the implication is that the type and style of a scene play a significant role in the scene description process by affecting the perceived subscene as well as the following SemRep formation process. The purpose of our experiments on this aspect is to show that the properties of the scene (e.g. thematic complexity, perceptual prominence of scene items, etc.) affect the immediate availability of a layout and the area of recognition. Although directly addressing these issues is very unlikely, we used some indirect methods, such as the semantics of an utterance, to measure the effects in the experiments. The idea is that if the layout of a scene is easily recognizable, a speaker is more likely to perceive the scene through a larger subscene,

eventually producing utterances describing a wide region of the scene (e.g. the theme of the scene), whereas a subject viewing a scene with an uneasily recognizable layout might end up describing a relatively smaller area of the scene (e.g. an object) since the scene is perceived via a smaller subscene.

Moreover, when we described the computational mechanisms of the scene description process within the framework of TCG, we also made a number of assumptions and hypotheses. Especially, we proposed a few principles that ground the production process of TCG (Section 4.5) – the premature production, the utterance continuity and the verbal guidance principles. The premature production addresses the situation where an utterance is made "before" the sentential structure is completely prepared or all of its constituents are figured out. A prematurely produced utterance may not always result in a fragmented or broken utterance since the utterance continuity principle allows the later produced utterance to be in grammatical continuity with the earlier produced utterance (even with some pause in between). Moreover, the verbal guidance principle covers the case where a prematurely produced utterance biases the visual attention process as the order of attention for identifying missing constituents is biased by the order of production of those constituents – e.g. the object corresponding to the subject of the sentence is more likely to be attended first.

These principles are interconnected around the notion of the threshold of utterance, which we defined as an upper bound on the available computational resources for producing sentences in TCG. Assuming that the premature production principle is already in effect, low threshold may cause an utterance to be produced prematurely, setting the stage for the utterance continuity and the verbal guidance principles to come into play. Thus, threshold is one of the key theoretical constructs of the production process of TCG and the interplay of the principles with threshold is tightly related to different degrees of the "well-formedness" of a produced utterance. Low threshold is proposed to be generally associated with production of more fragmented utterances while high threshold yields more complete sentences.

Thus, our focus is to show experimental support for the threshold of an utterance and assess the effects of various levels of threshold in the patterns of subjects' perceptual and verbal responses. Among various factors that influence threshold (e.g. individual preference, scene complexity, task requirements, etc.), we used time pressure to manipulate threshold in the experiments, using limited time as the means to lower threshold while allowing longer time to raise threshold. By examining the well-formedness of an utterance in various aspects under a different level of time pressure, we demonstrated (although indirectly) that threshold acts as a significant factor in the process of scene description. Although the well-formedness of an utterance may address many different properties, we mainly focused on the structural aspects, such as grammatical complexity (e.g. sentential length, embedded clauses, etc.) and production fluency (e.g. pauses, mumblings, filler sounds, etc.). This is based on the idea that when threshold is too low, utterances may be produced prematurely with an incomplete sentential structure or unprepared constituents, resulting in shorter, grammatically simpler sentences produced less fluently.

Experiment 1 (Section 5.2) provides empirical evidence for threshold based on eye-tracking data. In order to inflict a difference in time pressure, we used two independent tasks, the online and the offline, in which time pressure is applied in an all or nothing manner – the former requires subjects to describe a scene quickly while viewing a photo, whereas the latter requires subjects to inspect a scene first and then describe the scene after its image disappears. We provide results on simple measurement on subjects' utterance to highlight the effect of threshold supposedly imposed by the two different tasks, whose assessment is extended subsequently to Experiment 2. Another focus of Experiment 1 is to empirically address the validity of

the proposed principles of utterance production in TCG. Analysis result on the effect of the properties (or types) of the scene is also reported.

In Experiment 2 (Section 5.3), the difference in time pressure is elicited by two different requirements, in which subjects are asked either to describe a scene as quickly as possible or to describe the scene by taking as much time as needed. Both tasks are online as subjects produce the description while watching the scene. The collected utterance data were analyzed by measuring two factors, structural compactness and grammatical complexity. The former basically measures how compact the produced utterances are in terms of the ratio of the number of sentential structures to produce content words while the latter measures the complexity of grammatical structure. The effect of the different time pressures on these factors was measured. In addition, the effect of the scene style is also discussed in more detail than in Experiment 1. Subjects' initial utterances and gaze fixations before the utterance onset were analyzed in an effort to provide objective evidence for the proposed mechanisms for perceiving a subscene and the successive formation of SemRep.

Moreover, different levels of threshold, combined with the subscene perception process as specified above, may result in various patterns of gaze fixation and utterance production. The interplay between the resource constraints imposed by threshold and the layout availability of the perceived event may drive the system to generate different cases of scene description, even including some extreme ones – the system produces a highly time-locked pattern between fixation and utterance in certain cases while it produces utterances that bear the least correlation with the order of fixations in other cases.

In fact, there are two opposing views in the literature with regard to scene perception and speech production as discussed in detail in Section 5.4. One of these views claims that holistic conceptual structure governs the language production process (the structural view) whereas the other claims that the order of perceptual and conceptual input directly influences the linguistic output (the incremental view). We demonstrate these two views highlight two extreme patterns resulted from certain combinations of threshold and subscene. The idea is that one can elicit different eye gaze and utterance patterns through manipulating threshold (by imposing different levels of time pressure) and the coverage of a perceived subscene (by using scenes with different perceptual and conceptual properties). Although it is generally implausible to categorize a real case into a certain pattern, the specific conditions to result in extreme patterns within a limited circumstance are discussed in detail in Section 5.5.

5.2. Experiment 1

In our experiments, we used complex and natural scenes for visual stimuli with basically no restrictions on the syntactic format or style of speakers' utterance (except for the constraints naturally imposed by the experimental tasks). This is to explore various fixation and utterance patterns and observations that are not easily discerned from the highly constrained experimental settings used in most earlier experiments – controlled line drawings of simplified scenes were used as stimuli and the produced utterances were restricted to a certain simple sentence structure (e.g. Gleitman, et al., 2007; Griffin, 2004; Griffin & Bock, 2000; Meyer, 2004; van der Meulen, 2003).

Unlike those studies, our experiment mainly concerns the global-level relation of eye movements and utterances during scene description. Especially, the conventional type of statistical analysis is less helpful due to the large variation between individuals in this type of experiment with natural settings. Thus, the analysis of experiment data requires a specialized tool,

and we have designed new analysis software, *EyeParser*, for this specific purpose. Upon receiving raw eye-trace data and transcribed speech (manually transcribed with time stamps and some annotations), EyeParser generates various types of representations, such as a time-lined chart, text table, fixation distribution map, or a real-time movie clip.

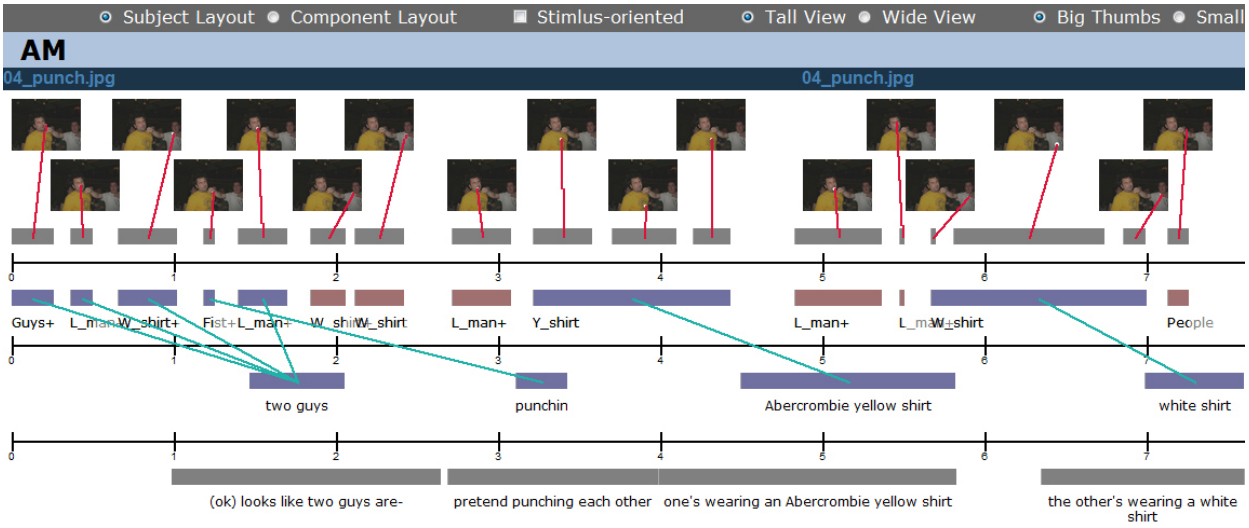


Figure 5.2-1: A screenshot of an EyeChart generated by EyeParser. Four types of data are represented along the time line: (1) the first row shows eye fixations superimposed on the stimulus image, (2) the labeled names of the areas around each fixation are represented in the second row, (3) key words (e.g. names of scene items) of utterances are linked with the matching fixations in the third row, and (4) the fourth row displays the entire transcribed utterances.

Among others, the time-line chart, which we call *EyeChart*, displays superimposed eye fixations with transcribed utterances from one or more subjects, which are charted along the time line (see Figure 5.2-1 for a screenshot of EyeChart from one data set). Although there are several other similar types of charts available (e.g. the multimodal time-coded score sheet by Holsanova, 2008), EyeChart enables the experimenter to cross-compare experiment data between subjects, visual stimuli or data types in relation to the passage of time, thus providing a "big picture" of the result. Moreover, an EyeChart is produced as a regular HTML file that can be read by any web browser and accessed in a highly interactive way – e.g. the experimenter can rearrange items in real-time, scroll the chart, or position the mouse cursor to enlarge thumbnail images and texts.

In addition, we thank Brenda Yang, a former undergraduate student of USC, for her help in hiring subjects and transcribing recorded speeches.

Now we provide a detailed account of the conditions and results for the experiment. Sample EyeCharts for all of the scenes used in this experiment are provided in Appendices.

A. Participants

Eight native or quasi-native English-speakers with normal or corrected-to-normal vision participated for course credit or complementary cash of $10. They were all undergraduate students of the University of Southern California (USC).

B. Visual Stimuli

We used photographs of natural indoor or outdoor live-action scenes (in full color) which include (but are not limited to) people making explicit transitive actions or implicit interactions between them. Each of these scenes captures a snapshot of an interesting and complex situation in which multiple events or aspects of an event at a certain moment are depicted (e.g. a guy is punching another guy while surrounded by a bunch of friends in a bar, who are laughing at the happening) in order to elicit variety of eye movement and utterance patterns. Only scenes which prominently included events happening between humans were selected.

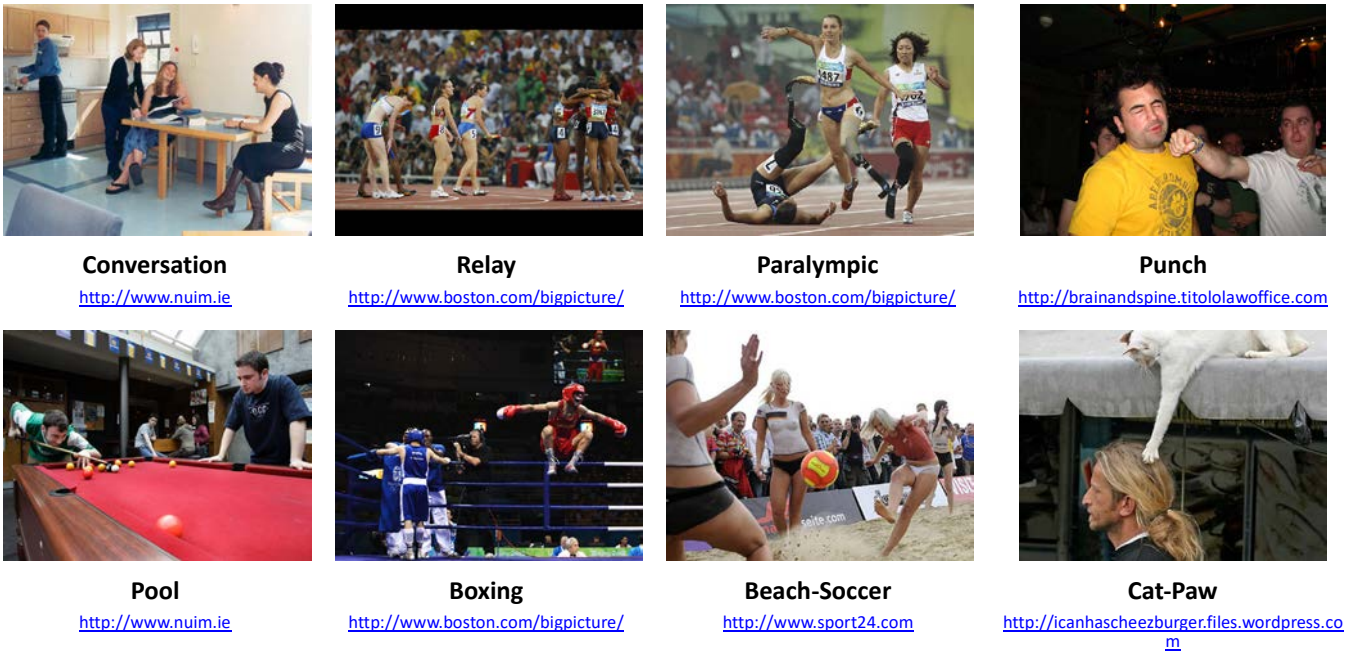


Figure 5.2-2: All the scenes used in Experiment 1. They are photographs of natural live-action events that are grouped into four categories as represented by each pair of the examples. The URL for the image sources are given under the names of scenes.

Eight scenes were chosen (except for practice scenes), which were then grouped into four types depending on the semantics and structure of the events in order to inspect the influence of the event structure to the patterns of eye movement and utterance. The scene types were specified as follows:

(1) actors of the scene casually interact each other with some sub-events happening in the background (e.g. people having a conversation in a kitchen),

(2) actors interact rather implicitly so that subjects need to infer what is going on (e.g. two groups of people gather together in a track meet),

(3) close inspection reveals another (rather shocking) fact which is not easily perceivable at a first glance (e.g. a race scene from the Paralympics),

(4) a very simple transitive action happens between actors (e.g. a guy punched by another guy).

Figure 5.2-2 illustrates all of the scenes used in this experiment. Each of the top-bottom pair of the illustrated scenes in the figure forms a category group – e.g. Conversation and Pool are in the category (1).

C. Apparatus

Visual stimuli were displayed on a 46-inch LCD monitor (Sony Bravia XBR-III, 1,016 × 571.5mm), 97.8cm in front of participants (corresponding field of view is 54:7° × 32:65°). The height of the seat was adjusted and a fixed helmet was worn over the participant's head to keep the position of the head fixed relative to the screen. The helmet was used instead of a chinrest in order to allow participants to move their mouths during scene description. Eye position was tracked by an ISCAN RK-464 (ISCAN) in pupil-CR mode (240Hz) to right eye following a 9-point calibration procedure. Participants' speech was recorded through a microphone placed in front of their mouths. We thank Dr. Laurent Itti (and the members of iLab) of USC for the use of this equipment and technical assistance.

D. Procedure

Before the experiment, it was explained to subjects that their eye movements and speech would be recorded, but nothing was specified about a form for the description or the types of scenes they would describe in order to elicit as natural responses as possible. Subjects were also asked to keep their head position still even if they were speaking. Subjects were asked to describe aspects of the displayed scene, but with different timing required depending on the tasks. The specifications are as follows:

- **Online Task:** Subjects were asked to describe the displayed scene *while viewing the photograph.* They were also asked to describe the scene *as quickly as possible*.
- **Offline Task:** Subjects were asked to describe the displayed scene *after it had disappeared*. There were no instructions on the speed of description.

The timing for display and description is set out below. The time course of fixation on and mentioning of objects were recorded for each trial.

Each subject completed a total of 10 trials (2 for practice, and 8 for stimulus scenes), and the two types of task were distributed randomly (uniform distribution) among the trials, 4 for each type. Before starting a trial, an instruction was displayed at the center of the screen – for the online task, "*Describe what you are seeing AS QUICKLY AS POSSIBLE*", and for the offline task, "*Describe what you HAVE SEEN AFTER the scene disappears*". The instruction remained on the screen until the experimenter clicked the mouse button, and then the trial began.

At the beginning of the experiment, subjects were properly positioned by installing the helmet over their heads with adjustment of the height of the seat. The eye-tracker was then calibrated while they fixated 15 successive positions across the screen as indicated by a blinking crosshair. Calibration typically took about 2 minutes. At the beginning of each of the experiment trials, after the instruction was displayed, a crosshair blinked for a brief period (1 sec.) at the center of the screen, followed by the display of a scene. For the online task, a scene was displayed for 15 sec., and during this time, subjects described the scene. For the offline task, a scene was displayed for 10 sec. while subjects refrained from speaking. After that, the scene disappeared and a blank screen was shown. For the online task, it was for 5 sec., and subjects may continue on the description if they did not finish although the speech recorded during this period was omitted from analysis. For the offline task, a blank screen was shown for 10 sec., and during this time, subjects described the scene from their memory. The instruction for the next trial appeared after the blank screen display.

E. Data Analysis and Results

For analysis, two types of data were collected: raw eye-trace data and recorded speech. Eye-trace data consisted of the horizontal and vertical position of the eye, which after a preliminary analysis were tagged with the status information (e.g. whether it is a saccade or fixation). Subjects' speech was recorded as audio bit-stream files, which later were transcribed manually into a text format with time-stamped and annotated words and phrases by using a standard audio file player – we used Audacity 2.0 (http://audacity.sourceforge.net/) for transcription. The experimental results provided in this section were firstly processed by EyeParser (mostly with EyeCharts) and then went through more specific analyses according to the type of measurement.

Utterance Well-formedness

As noted earlier, the effect of different levels of threshold on the produced utterance is one of our main concerns in this experiment. More specifically, our analysis focuses on whether or not low threshold caused the production of premature utterances. In fact we found that, even when subjects produced utterances for the same scene under the same task and their utterances represent very similar semantics, the sentential structure and the grammatical competence of their utterances may vary significantly. The followings are two example utterances produced by the subjects AM and TH during the online task for the same scene (the Relay scene in Figure 5.2-2).

Subject: ER

```
uh... this...
uh... looks like a-...
at track meet or something
uh... uh... a relay race
it's a bunch of um...
runners standing together
after a race
there are two teams
one has won and one has...
lost
```

Subject: TH

```
um there're
runners again
uh...
in the baton pass
there's a team of black women
who're huddling together
seem like they
uh...
have...
won a match
the other women
seem very upset
most likely that they lost
```

Their semantics are very similar – they both describe the entire theme of the scene first, and then move to the group on the right side, and then to the left group. However, their grammatical competence, or well-formedness, is quite different. ER's utterances were relatively fragmented (shorter sentences and more pauses) and grammatically inappropriate (missing subjects or unfinished sentences), whereas TH's utterances were comparatively "intact" in such a way that the sentences were longer with more complex structures, such as embedded clauses. Therefore, subjects' utterances indeed exhibit various levels of well-formedness, and we posit threshold as such a property that determines how much grammatical competence an utterance would exhibit.

Although the subsequent experiment described in Section 5.3 addresses it in further detail, here we briefly inspect whether or not the scene description process can be influenced by certain factors. Among such factors, we chose task requirements, as specified by the online and offline tasks in this experiment, to manipulate the level of threshold and simply

compared the counted numbers of words, sentences, and pauses from the utterance recorded under the different task requirements. This is basically to measure the effect of time pressure induced by the task requirement on the structure of the produced utterances.

We define a sentence as a group of words that form a grammatical structure that corresponds to a sentence with syntactic consistency and appropriateness. Clauses connected with conjunctions are considered as a single sentence only when they form relational connections (e.g. causal, logical), and this excludes the case where two clauses connected via a conjunction "*and*" without any strong relationship. Similarly, simple named object names are counted as separate sentences as they lack relational connections. Moreover, unfinished sentences (e.g. due to the change of plan) are excluded from counting. The following table (Table 5.2-1) summarizes a few example cases of counted sentence for analysis.

Table 5.2-1: Examples of counted sentences from different styles of utterance.

Utterance	Sentences
the man's falling down... while the other's kicking him	**1 sentence** – two clauses are combined into a coherent sentence by the conjunction *while*.
there's a bright yellow car and there's photographers taking pictures	**2 sentences** – two clauses are simply uttered without any relational connection.
there's a person who's fallen down and she is... they're handicapped	**2 sentences** – the unfinished clause *she is…* is excluded from counting.
uh... a woman in white dress a man in white tux	**2 sentences** – simple consecutive phrases for the woman and man are counted separately.

Similarly, we specified the definition of a "pause" for our own purpose – a pause is defined as an uncontinous delay within a sentence structure that lasts at least 300ms. This means breaks between sentences are *not* counted as pauses since we see them as a type of natural delay unrelated to the fluency of speech. Pauses include silent pauses, prolonged sounds (e.g. "*people are*…"), and verbalized pauses (e.g. "*uh*…", "*um*…"), and they are represented as three consecutive dots ("…") in the utterance examples of ER and TH.

As mentioned earlier, we are interested in the effect of threshold imposed by the type of task on the degree of well-formedness of the produced utterance. Our measure is defined as ratios between three factors of subjects' utterance – the number of words, sentences, and pauses. We used the ratios because simple comparisons between the numbers of words or sentences is not appropriate mainly due to the different time duration between the online and offline task – scene description time for the online task was 15 sec. and that for the offline task was 10 sec.

The number of words was simply measured by counting all grammatically appropriate words, except for the words in unfinished sentences since those sentences were also excluded from counting sentences. The numbers of other factors were measured as specified earlier. The ratio between words and sentences, the *word-sentence ratio*, which was calculated by the number of words divided by the number of sentences, is expected to provide an estimate on how structurally compact the produced utterance is (i.e. more words per sentence) while the ratio between pauses and sentences, the *word-pause ratio*, which was calculated by the number of pauses divided by the number of words, is expected to provide an estimate on how

fluently the utterance is produced (i.e. more words per pause). For example, the word-sentence ratio of ER's utterance (32 words, 7 sentences, 8 pauses) was 4.57 and the word-pause ratio was 4.0, whereas the word-sentence ratio of TH's utterance (34 words, 4 sentences, 3 pauses) was 8.5 and the word-pause ratio was 11.33 – according to this measurement, TH's utterance is highly more well-formed than ER's.

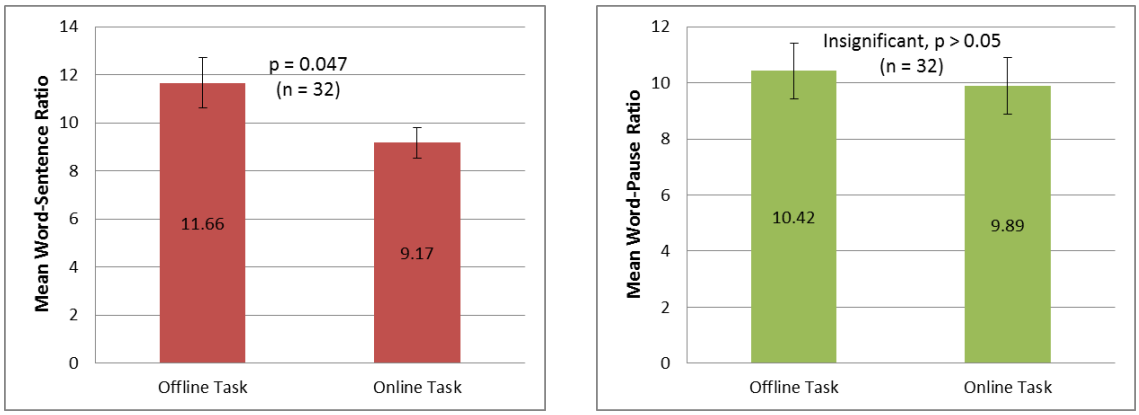


Figure 5.2-3: The mean word-sentence ratio (left) and the word-pause ratio (right) measured from subjects' utterance. Although both ratios were higher for the offline case, only the word-sentence ratio showed a (marginally) significant difference depending on the task type.

As shown in Figure 5.2-3, the result showed that the word-sentence ratio for the offline task was significantly higher than that for the online task (the mean ratio was 11.67 for the offline task and 9.17 for the online task, t-test, $p = 0.047$). This implies that subjects generally made longer sentences during the offline task, and this supports our hypothesis that a high threshold allows the system to produce more well-formed utterances since the sentence length is associated with the grammatical complexity and sentential formality. We also tried to analyze the ratios in the subject-wise, but none of the result showed a significant difference, mainly due to the lack of the number of samples – each subject produced only 4 utterances for each task type (8 utterances total) and this is too few to draw any statistical conclusion.

On the other hand, analysis on the word-pause ratio did not show any significant difference between the offline and online task. Although the mean ratio was higher for the offline task, the comparison between the two task types did not yield statistically significant different word-pause ratios (the mean ratio was 10.42 for the offline task and 9.89 for the online task, t-test, $p = 0.68$). Thus, the implication is that subjects produced as many pauses during the offline task as they did during the online task, and threshold may not be a significant factor in determining the frequency of pause – it may be a personal characteristic of each individual speaker. In fact, Ferreira and Swets (2002) reported that compared to the case when they were instructed to speak as quickly as possible, subject showed more "incrementality" (more inter-word delays) in producing utterances when a drastic means for imposing time pressure was taken – a timing bar was displayed on the screen, and if utterance was not made before the timing bar counted all the way down, a loud "beep" sound was produced. Thus, applying such a measure may produce a significant difference in subjects' frequency in making pauses.

Although the results were limited, the indication is that the produced sentences were influenced by the task type to some extent – longer sentences were produced during the offline cases, which are associated with high threshold, while online cases, which are associated with low threshold, yielded shorter sentences. More analysis on this aspect is given in the

subsequent experiment in Section 5.3.

Utterance Production Principles

Among the hypothesized principles of TCG, the principles that are especially in close relation with the notion of threshold are premature production and utterance continuity. Premature production addresses the case where an utterance is made before the sentential structure is completely prepared, and it is generally caused by too low a threshold. The utterance continuity principle allows the prematurely produced utterance to be smoothly connected with the successive utterance with an appropriate grammatical integrity.

Although we did not find a strong tendency in subjects to produce more pauses and breaks during the online task compared to the offline task, we found that they *did* produce reliably frequent pauses during production – 4.52 pauses per trial, the standard error of the mean (SEM) = 0.20, and one pause in every 10.15 words, SEM = 0.59. However, despite these pauses, the frequency of utterance incontinuity was pretty low (0.35 incontinuities per trial, SEM = 0.076) and only 7.7% of the pauses (22 out of a total 285 pauses) resulted in grammatically incongruent sentences – most of the time, subjects produced a gerundial or relative clause for appending a new clause. Thus, these results suggest that principles of the premature production and utterance continuity, or any other similar principles, may come into play when subjects produce utterances – during the scene description task, they indeed produced pauses and breaks intermittently while they managed to maintain the grammatical congruency between broken utterances.

Moreover, the last principle that we hypothesized for the production process of TCG is the verbal guidance, which addresses the case where the utterance structure under formulation guides visual attention in the way that an object mentioned next is more likely to be attended first among other objects even if it is perceptually less salient. Direct validation of this effect in real situations is highly improbable since it is difficult to judge whether an object is attended first because it will be mentioned next or it is mentioned first because it has been already attended. Thus, we used an indirect approach in the analysis, in which we measured the duration of fixations made during the scene display period and compared the mean durations in different task types. Since the online task required subjects to produce utterances while watching a scene whereas the offline task allowed subjects to focus on examining a scene without any verbal interference, we expect the mean fixation duration measured from the online task to be different from that of the offline task. The basic idea is in line with the finding of Papafragou and colleagues (2008) where speakers' eye movements generated significant cross-language differences during a verbal description task, which is in contrast to a free-viewing task where no such language-dependent differences were found.

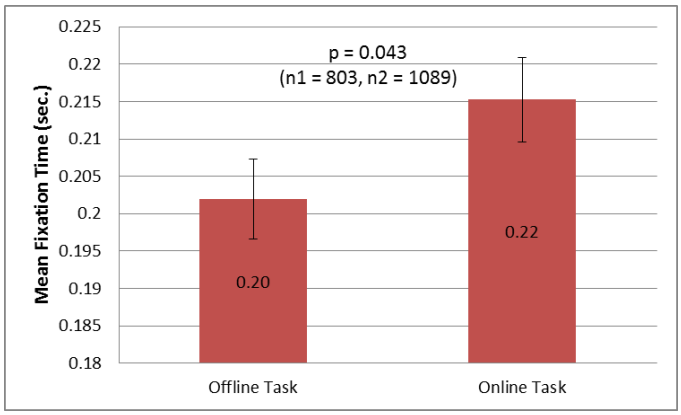


Figure 5.2-4: The mean fixation time measured in comparison between the online and the offline task, indicating that subjects fixated longer at objects during the online task. Since scene display time for the online task was longer (15 sec.) than the offline task (10 sec.), subjects generated more fixations during the online task (1089 fixations for total) than the offline task (803 fixations for total).

As we expected, there was a statistically significant difference in the durations of subjects' fixation time depending on the task type (the mean fixation time for the offline task was 0.20sec. while it was 0.22sec for the online task, t-test, $p = 0.043$) – they generally fixated longer at objects in the scene during the online task. The task requirements that made subjects produce verbal expressions during visual inspection seems to be the reason for the longer fixation time; in addition to the time for identifying an object, subjects may need to fixate on the object longer for retrieving the verbal information, such as phonological form of object names (Meyer, Sleiderink, & Levelt, 1998). In fact, speakers were reported to view an object more than twice as long when complex noun phrases were produced (e.g. *the big red scooter*) compared to simple noun phrases (e.g. *the scooter*), indicating that subjects' eyes remain on the referent object until they have fully planned the phrase to the point of initiating the phrase-final word (Levelt & Meyer, 2000; Meyer, 2004). Similarly, it has been reported that speakers allocated visual attention both less often and for shorter periods to objects for pronouns than for full noun phrases (van der Meulen, Meyer, & Levelt, 2001).

Although the evidence presented here is indirect and limited, the indication is that verbal requirements can bias the process of visual attention and eye movement, supporting the verbal guidance principle – the requirement of producing a verbal description during the online task might have biased the vision processes to generate longer fixations on the perceived objects, possibly for retrieving the verbal information in addition to the basic identification of the objects.

Subscene and Scene Perception

One of the most important concepts proposed for the scene perception process within the framework of SemRep and TCG is the notion of subscene and the relevant visual perception mechanisms. It is proposed that a scene is perceived in terms of subscenes, which capture cognitively significant aspects of a scene, and the system forms a SemRep in accordance with the events and entities delineated by those perceived subscenes. The key point in perceiving a subscene and the formation of a SemRep is that the process can be executed through different procedural steps depending on the perceptual and conceptual properties of a scene – a scene with a difficult event may be perceived by extending a subscene incrementally

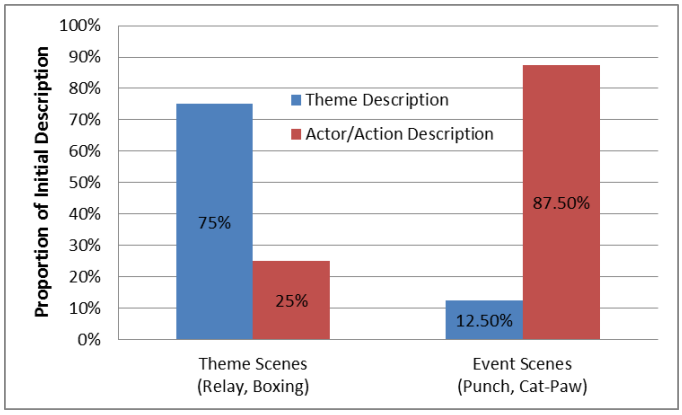
while a scene with an easy event may be perceived by a wide subscene covering the entire event, whose details are filled in subsequently. The area of coverage and the level of detail of an immediately perceived subscene are the determinant factors of the proposed scene perception process, and these factors are proposed to be affected by the properties of a viewed scene.

Although the subsequent experiment described in Section 5.3 addresses in further detail, our analysis focus here is to briefly assess the influence of the properties of a scene, such as event difficulty, to the perception process. Since the direct measurement of either the size of subscene or its level of details is not possible, we inspect the initial utterances of subjects as an indirect method. The underlying idea is that if the layout of a scene is easily recognizable, a speaker perceives the scene more likely through a larger subscene, possibly producing utterances describing a wide area of the scene (e.g. the theme of the scene), whereas a subject viewing a scene with an uneasily recognizable layout might end up describing a relatively smaller area of the scene (e.g. an action or actors of the scene) as the scene is perceived via a smaller subscene.

We chose two groups of scenes in this analysis: one for events with ambiguous relationships between actors, which are not easily recognizable (*theme scenes*), and the other for events with very clear and simple actions (*event scenes*). The Relay and Boxing scene were selected for the former group and the Punch and Cat-Paw scene were selected for the latter. Subjects' initial utterances were also analyzed into two categories, theme description and actor/action description, depending on the semantic coverage and the level of description details. Utterances corresponding to the *theme description* generally describe the theme of the entire scene without mentioning much detail whereas utterances corresponding to *the actor/action description* generally focus on certain aspects of the depicted event in a relatively detailed manner. Specific examples are summarized in the following table.

Table 5.2-2: Example initial utterances of different description types.

Initial Utterance	Description Type
`oh wow we have a cage fighting match`	Theme description
`in this scene it looks like it's a boxing match`	Theme description
`the cat touching a man's head`	Actor/action description
`(ok) looks like two guys are... pretend punching each other`	Actor/action description



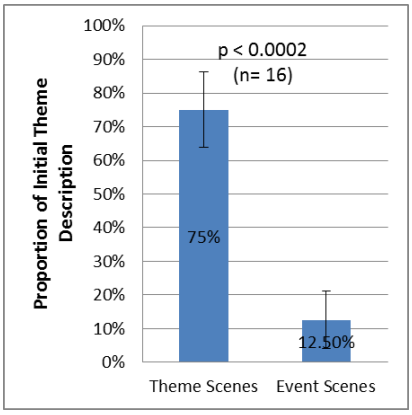


Figure 5.2-5: The proportion of description type analyzed from subjects' initial utterances – both of the theme and actor/action description type represented for each scene type (left), and only the theme description type was compared (right). Subjects generated initial utterances of a significantly different description style depending on the scene type.

Since we chose two scenes for each scene type and the total number of subjects was eight, 16 initial utterances were analyzed. The result indicated that subjects produced more utterances corresponding to the theme description for theme scenes while more utterances corresponding to the actor/action description for event scenes (75% of the total initial utterances were of the theme description for theme scenes, and 87.5% of utterances were of the actor/action description for event scenes). As hypothesized, a clear tendency of subjects to produce initial utterances that match with the scene type was observed, and the observed tendency was statistically significant (for theme scenes, 75% of the total initial utterances were of theme description, SEM = 11.1%, while only 12.5% of the utterances were of theme description for event scenes, SEM = 8.5%, t-test, $p = 0.00013$).

One particular thing to note about this analysis is that the utterances from both of the tasks were analyzed altogether and the effect was found to be significant. In fact, separate analyses on the utterances from the online and offline tasks yielded significant tendencies (analysis on the online and offline tasks resulted in exactly the same result as 75% of the total initial utterances were of theme description and 12.5% of the utterances were of theme description for event scenes, t-test, $p = 0.0095$). Thus, the initial utterances recorded not only from the online task but also from offline task yielded significantly different description coverage. The implication of the result is two-fold: (1) the effect of scene properties was strong enough to last over a relatively long period (10 sec.), and/or (2) the formation of the sentential structure was done in a deterministic way such that the selection made for the earlier utterance was not influenced by the selection of the subsequent utterances. In any case, the influence of the scene properties seemed to be significantly strong in formulating utterances.

Although we took an indirect measurement, the implication of the results is that an initially perceived subscene may be affected by the conceptual and perceptual properties of a watched scene as illustrated by the different description coverage and detail of subjects' initial utterances induced by different scene types.

An experiment with a similar paradigm was conducted previously by Fei-Fei and her colleagues (2007). They showed gray-scale photographs of natural scene to subjects and asked them to produce descriptions while randomly varying the presentation time (from 27 to 500ms). They analyzed the produced descriptions by assigning individual scores to different attributes appeared in the description (e.g. indoor/outdoor, object animacy, event types, etc.). In contrast to our claim, they argued that there is little evidence for bias toward either scene-level or object-level recognition. However, despite the similarity in the approach, the implication of their findings is intrinsically incompatible with ours since our analysis focused only on the perception of an immediate subscene as reflected in the initial utterance while their analysis covered the entire description.

5.3. Experiment 2

Experiment 2 differs from Experiment 1 in a few ways. Firstly, we focus on the threshold of utterance and the perception of subscene, and provide analysis results in further detail. Secondly, the difference in time pressure was again elicited by two

types of task but with different requirements. In this experiment, subjects are asked to describe a scene either as quickly as possible or while viewing the scene for as much time as needed. Both tasks are "online" in that subjects produced a description while watching a scene. Thirdly, only two types of scene were used (but the total number of the scenes, eight, was identical) to allow simpler inspection of the effect of scene properties. Depending on the thematic and perceptual characteristics of the depicted events, scenes were categorized as either event or theme, namely one for events with ambiguous relationships between actors, which are not easily recognizable, and the other for events with very clear and simple actions, as discussed earlier in Experiment 1. Lastly, more subjects participated in the experiment. It was not only convenient for statistical analysis as more data samples are available but also helpful for providing a variety in subject responses.

Now we provide a detailed account of the conditions and results for the experiment. Sample EyeCharts for all of the scenes used in this experiment are provided in Appendices.

A. Participants

15 native or quasi-native English-speakers with normal or corrected-to-normal vision participated for course credit or complementary cash of $10. They were all undergraduate students of the University of Southern California (USC).

B. Visual Stimuli

Identical to Experiment 1, full-color photographs of complex, natural live-action scenes were used while some of the scenes in Experiment 1 were used again in this experiment (with some retouches). Again, only the scenes with interesting and complex situations in which multiple events or aspects of an event at a certain moment are depicted were chosen.

Figure 5.3-1: All the scenes used in Experiment 2. They are grouped into two categories – the top row shows event scenes and the bottom row shows theme scenes. The URL for the image sources are given under the names of scenes.

A total of eight scenes were used (in addition to practice scenes), which were grouped into two different types depending on the thematic and perceptual characteristics of the depicted events.

(1) **A "theme" scene**: the "theme" is more prominent than each individual's action or interactions between them – the overall atmosphere or layout of the scene is more salient and prominent.

(2) **An "event" scene**: the "event" is more prominent than the theme of the scene – the action of each individual or the interaction between individuals is more salient and prominent.

Figure 5.3-1 shows all of the scenes used in this experiment. The top row contains event scenes (Punch, Bull-Run, Cholitas, and Kisses) while the bottom row contains theme scenes (Boxing, Car-Expo, Wedding, and Soccer).

C. Apparatus

The same experiment settings as Experiment 1 have been used.

D. Procedure

As in Experiment 1, it was explained to subjects that their eye movements and speech would be recorded, but nothing was specified about a form for the description or the types of scenes they would describe. Subjects were also asked to keep their head position still even if they were speaking. Similarly to Experiment 1, subjects were asked to describe aspects of the displayed scene, but with different time pressure imposed by instructions in an effort to manipulate their threshold of utterance. Two task types were specified as follows:

- **Quick Task:** Subjects were asked to describe the displayed scene *as quickly as possible* while viewing the photograph.
- **Free Task:** Subjects were asked to describe the displayed scene *taking as much time as they needed* while viewing the photograph.

Both tasks were "online" in the sense that subjects described the scene as they watched it. This was to enable inspection of the temporal correlation between eye movements and utterances. The quick task is to elicit a low threshold whereas the free task is to elicit a high threshold, and time pressure was imposed by specifically instructing subjects of the speed requirement.

Each subject completed a total of 12 trials (4 for practice, and 8 for stimulus scenes), and the two types of task were distributed randomly (uniform distribution) among the trials, 4 for each type. Before starting a trial, an instruction was displayed at the center of the screen – for the quick task, "*Describe what you are seeing AS QUICKLY AS POSSIBLE*", and for the free task, "*Describe what you are seeing while TAKING AS MUCH TIME AS YOU NEED*". The instruction remained on the screen until the experimenter clicked the mouse button, and then the trial began.

In both tasks, subjects were allowed to keep describing as much as they wanted (i.e. there was no time constraint), but they were advised to attend to most highlighted events or aspects of a scene rather than exhaustive details. The scene was displayed until the experimenter clicked the mouse button after subjects notified that the description had been finished. The time course of fixation on and mentioning of people and objects were recorded for each trial.

Other experiment procedures were identical to Experiment 1.

E. Data Analysis and Results

Identically to Experiment 1, eye-trace data and transcribed speech were used for analysis. These data were initially

processed by EyeParser (mostly with EyeCharts) and then went through more specific analyses according to the type of measurement.

Scene Type

Before we begin, we should address the details of a brief survey on the categorization of the scenes used in this experiment. Since the scene types are an important variable along with the task types, we conducted a quick poll to estimate whether the scenes are fitted with the already set up categories in an objective manner. 19 participants volunteered in the survey, all of whom were graduate students of USC. We asked participants to rate each of the scenes according to a scale that marks the characteristics of the depicted events with a numerical value that ranges from 1 to 5 (1: very event, 2: event, 3: hard to say, 4: theme, 5: very theme).

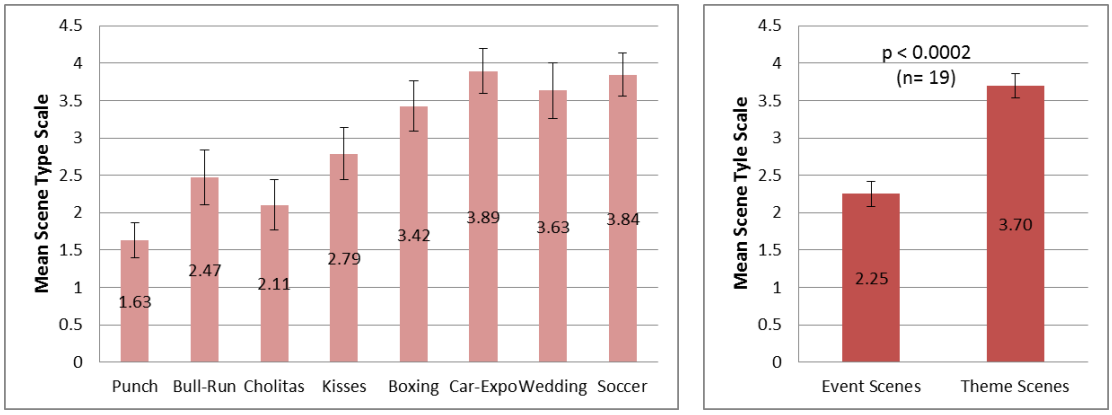


Figure 5.3-2: The mean scale value calculated from raters' responses for each scene (left) and collapsed for all event and theme scenes (right). The mean scale value for event scenes was significantly distinctive from the mean value for theme scenes.

As Figure 5.3-2 shows, participants' response generally conformed to the proposed division – all event scenes (Punch, Bull-Run, Cholitas, and Kisses) were rated on average below 3.0, which is the median of the scale, while all theme scenes (Boxing, Car-Expo, Wedding, and Soccer) rated above 3.0 on average. In fact, the mean scale value of event scenes was significantly lower than the mean value of theme scenes, validating the classification of scenes in the current experiment (the mean scale value was 2.25 for event scenes, SEM = 0.17, and 3.70 for theme scenes, SEM = 0.16, t-test, $p < 10^{-8}$). Moreover, Table 5.3-1 represents the results of a cross-comparison of all scenes, which exhibited the boundary of division that generally fitted with the current grouping of scenes. The mean scale values of the scenes in different groups were significantly different while those of the scenes among the same group did not show a significant difference.

Table 5.3-1: A cross-comparison of all scenes used in the experiment. The matching between two scenes is marked with an "X" if the difference in the mean scale values is insignificant whereas it is marked with an "O" if the values are significantly different (t-test, $p < 0.05$). The shaded cells represent the comparisons between the scenes of the different categories.

	Punch	Cholitas	Bull-Run	Kisses	Boxing	Wedding	Car-Expo	Soccer
Punch	-	X	X	O	O	O	O	O
Cholitas	-	-	X	X	O	O	O	O
Bull-Run	-	-	-	X	X	X	O	O
Kisses	-	-	-	-	X	X	O	O
Boxing	-	-	-	-	-	X	X	X
Wedding	-	-	-	-	-	-	X	X
Car-Expo	-	-	-	-	-	-	-	X
Soccer	-	-	-	-	-	-	-	-

Effect of Time Pressure

Differently from Experiment 1, the current experiment elicited the effect of threshold by giving subjects instructions that presumably caused different levels of time pressure. Since the effect of threshold is one of the key factors that this experiment addresses, it is crucial that subjects indeed followed the instructions specified by different task types. We analyzed the onset time of utterances produced by subjects to estimate the effectiveness of the task difference.

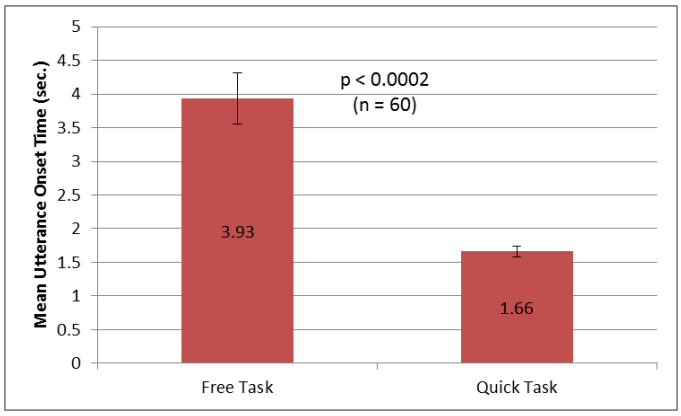


Figure 5.3-3: The measured mean utterance onset time during the free and quick task. Utterances were produced faster during the quick task, indicating that subjects were under more time pressure.

For each task type, 60 utterance data (15 subjects with 8 scenes divided by two tasks) were examined. The result showed a clear difference in the utterance onset time between the free and quick task. Subjects produced utterances significantly later during the free task than during the quick task (the mean onset time was 3.93 sec. for the free task, SEM = 0.38, and the 1.66 sec. for the quick task, SEM = 0.076, t-test, $p < 10^{-6}$), confirming that the given instructions were effective in yielding the expected time pressure.

Utterance Well-formedness

As discussed in Experiment 1, the effect of different levels of threshold on the produced utterance is one of our main issues. In fact, the analysis on threshold in Experiment 1 suggested that low threshold may cause the production of less well-formed utterances. Although threshold did not appear to have any effect on producing pauses or breaks during utterance, the sentence length did seem to be affected by threshold – the word-sentence ratio for the offline task was significantly higher than that for the online task. We suspected that a high threshold allows the system to produce more well-formed utterances as the sentence length is generally associated with the grammatical complexity and sentential formality. Compared to Experiment 1, especially, both of the tasks were designed as "online" in order to ensure higher credibility in the direct comparison of utterance structures from the two types of task.

In this analysis, we examined the structural aspect of the produced utterance more closely to estimate the extent to which threshold influences the well-formedness of utterances. There are a number of structural factors to be considered involved in the assessment of sentences with different degrees of well-formedness. For example, a scene where a woman who is wearing a blue dress is hitting a man can be described by very different styles of utterance, such as "*the pretty woman in blue hits the man*" or "*the woman hitting the man... she's pretty… and wearing a blue dress*". The former utterance may be the result of a high threshold where careful planning is possible with a higher working memory load and more constructions allowed, whereas the latter may result from a low threshold where the sentence formulation process happens with less computational resources available. Due to the insufficiency of the allowed computational resources, the utterances in the latter case may be produced in a highly incremental fashion, which generally ends up in grammatically simpler and shorter utterances, such as phrases. Thus, the difference between the former and the latter utterances, although their semantic meanings are almost identical, should be mostly reflected by the arrangement of semantic components and the complexity of the grammatical structure. In the former utterance, there is one big component, *the woman is hitting the man*, which "embeds" other subordinate components, *the woman being pretty* and *the woman wearing a blue dress*, inside. On the contrary, the latter utterance exhibits a plainer grammatical structure where the three components are simply concatenated one after another. The former is more well-formed than the latter as it conveys the meaning more compactly (more words in fewer sentences) with higher formality (more complex grammatical structures with embedded components).

To capture these aspects, we define two metrics, the *structural compactness* and the *grammatical complexity* of utterance, which are a more elaborate version of the metric (i.e. word-sentence ratio) used in Experiment 1 to provide an estimate on how structurally compact the produced utterance is. The formulas are specified as follows:

$$\textbf{Structural Compactness} = \frac{\textbf{Number of Core Words}}{\textbf{Number of Sentences}},$$

$$\textbf{Grammatical Complexity} = \frac{\textbf{Number of Embedded Structures}}{\textbf{Number of Sentences}}.$$

Core words are basically content words that are closely related to the semantics of the scene and objects being described. Not all content words in utterances are counted as core words while some function words (especially pronouns) might be since our main concern in this analysis is the measurement of how much of scene semantics, not all semantics in general, is

delivered through utterances. In fact, subjects often produced expressions that are not really meaningful in terms of the scene semantics – e.g. in the utterance "*so in this scene, it seems like there's a man*", only the word *man* is counted as a core word even though there are other content words, such as *scene* or *seem* (see Table 5.3-2 for more examples).

Embedded structures are clausal, prepositional, or other sentential structures that are structurally "embedded" within another structure. Embedded structures are of particular interest to us since they reflect the computational efforts (especially higher working memory overhead) of the production process of TCG. A high threshold allows the system to formulate utterances by populating multiple grammatical structures (in terms of constructions) at the same time, which generally result in sentential structures of higher grammatical complexity with multiple levels of relative clauses or other sentential structures. Not all relative clauses are counted as embedded because an embedded structure in our definition needs to appear "within" another structure – e.g. in the utterance "*a woman wearing a green dress is kicking a woman wearing a blue dress*", the relative clause *wearing a green dress* is counted as embedded while it is not in the utterance "*people are watching the woman wearing a green dress*" because it is possible that the clause is simply appended at the end of the main sentential structure rather than embedded within it (see

Table 5.3-3 for more examples).

The definition of a sentence is the same as the one described in Experiment 1 (see Table 5.2-1 for examples).

Table 5.3-2: Examples of counted core words from different styles of utterance.

Utterance	Core Words
it looks like there's a cameraman… filming… the man in blue on the side	cameraman, filming, man, blue, side
um looks like maybe… like a rare romantic… day I guess	rare, romantic
there are other two people in the background who… um… seem to be helping… with the event	other, two, people, background, help, event
this is taking place within a soccer game in which an opponent is… um… well accidentally attacked… uh his opponent	take place, soccer, opponent, accidentally, attacked, his, opponent

Table 5.3-3: Examples of counted embedded structures from different styles of utterance.

Utterance	Embedded Structures
one guy on the orange team just kicked a guy on the blue team in the chest	**2 structures** – orange team, blue team
all of his friends behind him are laughing	**1 structure** – behind him
the main man… who's the largest and the… the image is running	**1 structure** – the largest
there a racing model who's being employed to stand… next to one of the vehicles is being photographed by uh… onlookers	**1 structure** – employed to stand next to vehicle

Thus, the structural compactness and the grammatical complexity are supposed to provide simple numerical measures of the well-formedness of an utterance as the former measures how compactly represented the meaning of an utterance is while the latter measures how complex the grammatical structure of an utterance is. In the earlier example for the utterances "*the*

pretty woman in blue hits the man" and "*the woman hitting the man... she's pretty… and wearing a blue dress*", both metrics score much higher for the former (structural compactness = 5/1 = 5.0, grammatical complexity = 1/1 = 1.0) than for the latter (structural compactness = 8/3 = 2.67, grammatical complexity = 0/3 = 0.0). The following is an example analysis of the actual utterances from a subject.

Utterance from subject AT (for Punch scene)

```
oh
this could… potentially be like a joke
but
one man is… punching or making the action
like he's punching the other man
look-
looks like the man with the yellow shirt is
reacting to the punch
uh though there's a very enthusiastic man in
the green shirt in the background
who's laughing at the whole ord- ordeal
```

Analysis Result[4]

```
potentially, joke
------------------------------
one, man, punch, action, he, punch, other, man
------------------------------
man, [yellow, shirt], react, punch
------------------------------
enthusiastic, man, [green, shirt], background,
laugh, ordeal
```

Core Words = 22, Embedded Structures = 2, Sentences = 4

Structural Compactness = 5.5

Grammatical Complexity = 0.5

Two raters analyzed a total of 120 utterances (15 subjects for 8 scenes, 60 utterances for each task type) gathered from the experiment for the measurement of the structural compactness and the grammatical complexity. Then the metric scores from both of the raters were averaged to yield the final analysis result.

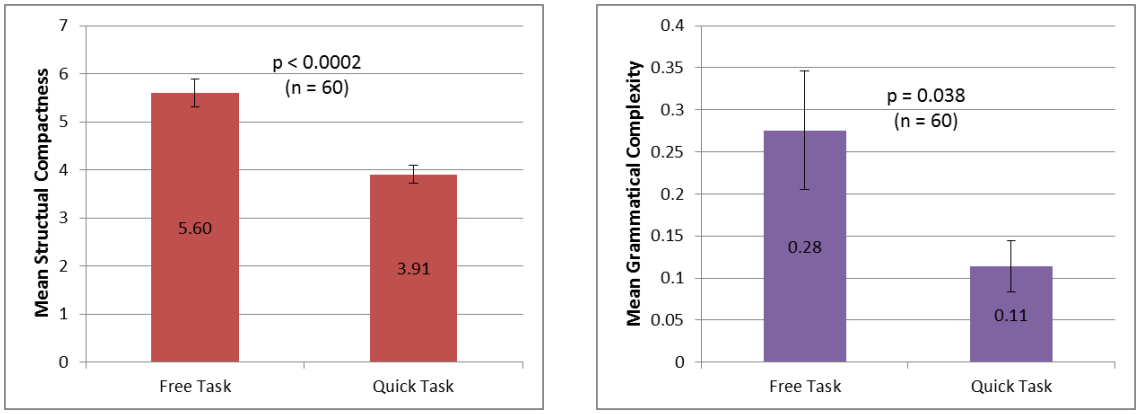


Figure 5.3-4: The mean score of the measured structural compactness (left) and the grammatical complexity (right). Both of the metrics resulted in a significantly higher score for the free task than for the quick task, indicating that the well-formedness of the produced utterances was influenced by threshold.

As shown in Figure 5.3-4, both of the measured metrics rated significantly higher for the free task than for the quick task

[4] Analysis was basically done by counting core words. The dashed line represents the sentence boundary, and the words within square brackets ("[ ]") represent the core words within an embedded structure.

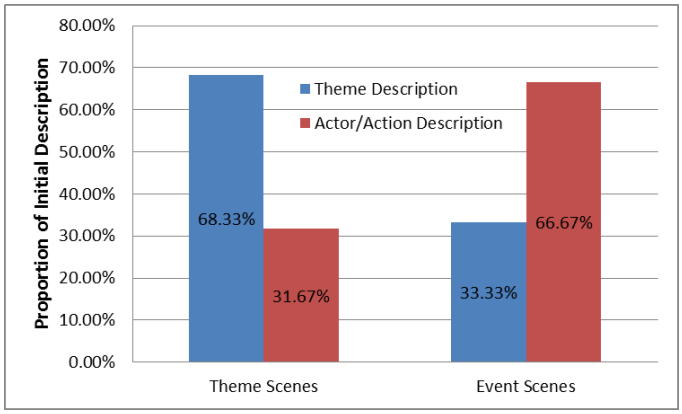
(for the free task, the mean score of the structural compactness was 5.60, SEM = 0.29, and the grammatical complexity was 0.28, SEM = 0.07, whereas for the quick task, the mean score of the structural compactness was 3.91, SEM = 0.19, and the grammatical complexity was 0.11, SEM = 0.03, t-test, $p < 10^{-5}$ for the structural compactness and p = 0.038 for the grammatical complexity). This indicates that the task with lower time pressure (i.e. the free task) results in more well-formed utterances, suggesting that the sentential structure of the produced utterance is influenced by the threshold level. The comparison between the mean score of the structural compactness for theme scenes and that for event scenes did not show any significant difference (4.92 for theme scenes and 4.59 for event scenes, t-test, p = 0.38), confirming that the difference was in fact caused by the task difference rather than the scene difference.

Moreover, compared to the structural compactness, the result of grammatical complexity was marginally significant ($p < 10^{-5}$ for the structural compactness, and p = 0.038 for the grammatical complexity). In fact, the scene-type-wise analysis for the structural compactness yielded a significant difference (for theme scenes, free task = 5.54, quick task = 4.2, p = 0.007, and for event scenes, free task = 5.7, quick task = 3.7, p = 0.0005) whereas the grammatical complexity for both of the scene types did not show any significant difference (for theme scenes, free task = 0.29, quick task = 0.11, p = 0.12, and for event scenes, free task = 0.26, quick task = 0.12, p = 0.17). Thus, the effect of threshold appears to be stronger in the structural compactness than in the grammatical complexity – i.e. subjects reliably spoke out longer sentences when threshold was higher, but the sentential structures were not necessarily more complex, especially with more embedded structures.

Subscene and Scene Perception

As addressed in Experiment 1, the notion of subscene with its related visual perception mechanisms is one of our main concerns along with the threshold. In fact, the results on this aspect in Experiment 1 suggested that the area of coverage and the level of detail of an immediately perceived subscene could be affected by conceptual and perceptual properties of a viewed scene.

In this analysis, we examined more closely the influence of two different types of scene, which were theme and event, in order to assess to what extent the scene properties induce the change in the level of coverage and detail of a perceived subscene. Identical to Experiment 1, we again analyzed the initial utterances produced from subjects (but with more number of subjects and scenes) in both task types. A total of 120 utterances (15 subjects for 8 scenes, 60 utterances for each task type) were analyzed.



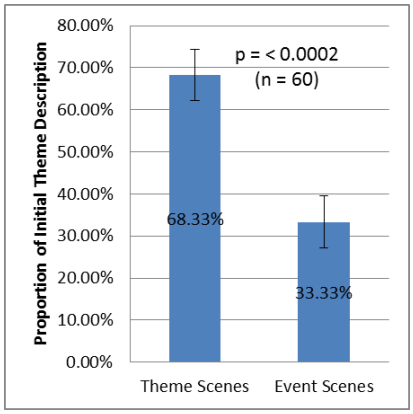


Figure 5.3-5: The proportion of description type analyzed from subjects' initial utterances – both of the theme and actor/action description type represented for each scene type (left), and only the theme description type was compared (right). Subjects generated initial utterances of a significantly different description style depending on the scene type.

As in Experiment 1, we found a clear tendency for subjects to produce more utterances corresponding to the theme description for theme scenes while more utterances corresponding to the actor/action description for event scenes (68.33% of the total initial utterances were of the theme description for theme scenes, and 66.67% of utterances were of the actor/action description for event scenes). The tendency was not as strong as what was found in Experiment 1 (75% for the theme description for theme scenes and 87.5% for the actor/action description for event scenes), possibly due to more variety in the actions and events depicted in the scenes used in the experiment – e.g. in Experiment 1, only the type of transitive action with the extension of an arm (a guy's punching and a cat's pawing) was depicted in event scenes whereas in Experiment 2, more various transitive actions, such as kissing and chasing, were used. However, the observed tendency was still statistically significant (for theme scenes, 68.33% of the total initial utterances were of theme description, SEM = 6.05%, while only 33.33% of the utterances were of theme description for event scenes, SEM = 6.1%, t-test, $p < 10^{-4}$).

Additionally, the tendency was significantly strong for both tasks. During the free task, 72.73% of the total initial utterances were about the theme for theme scenes while 44.44% of the initial utterances were about the theme for event scenes, and the difference was significant (t-test, $p = 0.028$). Similarly, during the quick task, 62.96% of the initial utterances were about the theme for theme scenes while 24.24% of the initial utterances were about the theme for event scenes, and the difference was also significant (t-test, $p = 0.0024$). The indication is that subjects produced more descriptions on the theme during the free task than the quick task (72.73% and 44.44% for the free task vs. 62.96% and 24.24% for the quick task) while the tendency of subjects to produce utterances of the matching style with the viewing scene was stronger during the quick task than the free task ($p = 0.028$ for the free task vs. $p = 0.0024$ for the quick task). It appears that during the free task, subjects had enough time to scan through a scene and plan the best sentential structure to describe the scene, and this might have resulted in greater production of thematic descriptions. On the other hand, the time constraint during the quick task enforced subjects to focus on the most salient aspect in the scene and immediately produce an utterance, and this might have yielded a more distinctive tendency in the styles of the produced utterances, which reflected the properties of the scene.

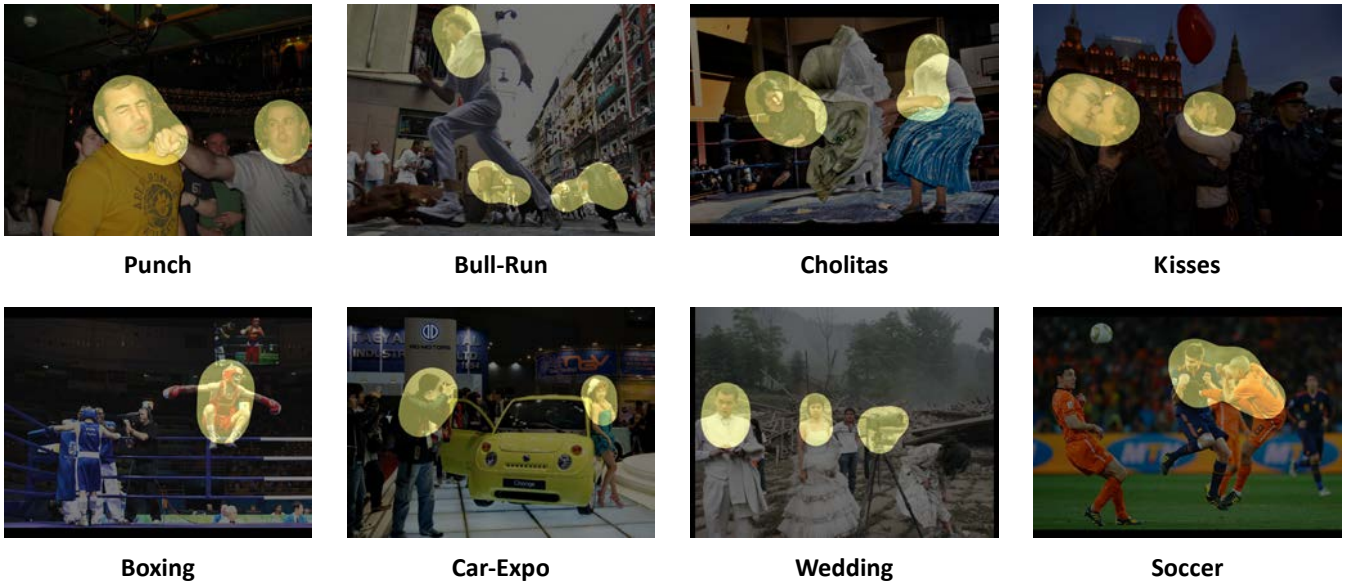


Figure 5.3-6: The marked regions of the actors/actions of the main event of each scene represented by the highlighted areas. These regions were selected according to the initial utterances of the actor/action description type.

In this experiment, moreover, we extended the previous version of analysis on the effect of the scene properties by measuring the locations of the gaze fixations made "before" the onset of the utterance. We define those fixations as initial fixations. Since the locations of speakers' eye gazes are tightly linked with the produced utterance (Levelt & Meyer, 2000; Meyer, 2004; Spivey, et al., 2004; van der Meulen, 2001) and even with typing (Andersson et al., 2006), initial fixations are expected to convey crucial information for forming initial utterances. Especially given the importance of the initial utterance in this analysis, close inspection in the initial fixations of subjects would be worthwhile in finding clues in the effect of scene properties and the formation of a subscene. We suspect that the locations of initial fixations would differ depending on the type of scenes as did the description coverage of the initial utterances presented earlier.

For measuring the difference in the locations of initial fixations, we marked all of the scenes used in the experiment with the regions of the actors and/or actions of the "main event". Figure 5.3-6 illustrates the marked regions of all scenes. The regions were selected by analyzing subjects' initial utterances for the description of actions or actors (i.e. the actor/action description) in terms of core words. Even for theme scenes, the actors and actions being described by subjects' initial utterances corresponding to the actor/action description were highly consistent such that only a few utterances described differently from the majority of the utterances, resulting in 89.8% of consistency overall (see Table 5.3-4 for more detail).

Table 5.3-4: The content and the proportion of agreement of the main event description appeared in the initial utterances produced for each scene.

	Punch	Cholitas	Bull-Run	Kisses	Boxing	Wedding	Car-Expo	Soccer
Core words	man, punch	women, fight, kick	man, run, chase, bull	couples, people, kiss	man, jump	couple, wedding photo	man, photo, girl, model	player, soccer
Majority Proportion	100% (9/9)	91.7% (11/12)	90% (9/10)	77.8% (7/9)	100% (6/6)	75% (3/4)	83.3% (5/6)	100% (3/3)

We then measured the proportion of the durations of the initial fixations that fell within the marked regions for each scene type – since the utterance onset time varied by subjects and cases, measuring and comparing the total or mean time of the fixation duration were inappropriate. Only the duration of fixations, which excludes the time for saccades, was counted into the proportion.

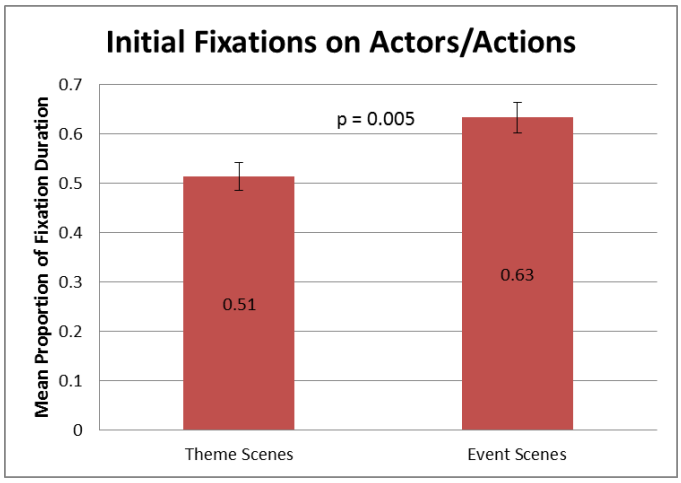


Figure 5.3-7: The mean proportion of the fixation duration time measured from subjects' initial fixations that fell within the marked regions of the actors and/or actions in the main event.

We found that subjects fixated on the actor and action regions significantly more for event scenes than for theme scenes (the mean proportion of the fixation duration time was 0.51, SEM = 0.028, for theme scenes and 0.63, SEM = 0.031, for event scenes, t-test, $p = 0.005$), confirming the effect of the properties of scene during the initial phase of scene perception suggested from the initial utterance analysis.

Interestingly, only the initial fixations during the free task yielded the desired effect with statistical significance as opposed to the quick task (during the free task, the mean proportion of the fixation duration time was 0.46, SEM = 0.036, for theme scenes and 0.59, SEM = 0.038, for event scenes, t-test, $p = 0.021$). It appears that during the quick task, subjects tended to fixate more on salient objects (mostly the actors/actions of the main event) even for theme scenes to extract enough information for the description as they were forced to produce an utterance as quickly as possible, and this might have resulted in a less significantly distinctive pattern of initial fixations. Combined with the previous result, which suggests that during the quick task, subjects' initial utterances showed a stronger tendency in matching the description style with the scene

type, the interpretation is that speakers form a subscene for the description of the theme of a perceived scene based on the "gist" rather than the detailed inspection on the entire scene with multiple fixations, especially when they are under time pressure. This is because when subjects produced utterances of the theme description during the quick task, they still tended to fixate more at the limited regions, suggesting that the information of the theme was acquired through some other means than thorough fixations. However, more research is needed to thoroughly address this issue.

Although the current finding does not directly address how a subscene is perceived or how much information is contained in a subscene at a first glance at a scene, an important implication is that the properties of a scene as well as the event of the scene actually affected the content of initial utterances and the patterns of fixations before those utterances. As addressed earlier, we propose that the area of coverage and the level of detail of an initially perceived subscene are reflected within those initially observed factors. Thus, the analysis results on initial utterances and fixation locations, in which we found a significant influence from the properties of a perceived scene, suggest that those properties influence the establishment of a subscene and the subsequent perception processes. However, more research is needed to fully address the validity of the scene perception mechanisms proposed in the current work (see Section 2.7 for the detailed exposition of the proposed process).

5.4. Two Views in Eye Movements and Speech Production

In recent years, many psycholinguists have used eye movements to study how speech and language are understood, to learn about the processes involved in language production, and even to shed light on how conversations are managed (Henderson & Ferreira, 2004). Given that the language model proposed in the present study accounts for production of speech, eye tracking studies on language production are of our main concern here.

Griffin and Bock (2000) monitored the eye movements of speakers while describing black-and-white line drawings of simple transitive events with single sentences. They found an orderly linkage between successive fixations in viewing and word order in speech. Especially, a similarity between speakers' initial eye movements and those of observers performing a nonverbal event-comprehension task was found, indicating that response-relevant information was rapidly extracted from scenes, allowing speakers to select grammatical subjects based on comprehended events rather than salience – i.e. people do not always start a sentence with what captures their eyes first. In fact, it has been reported that speakers do not readily give up a structural preference in order to put a perceptually salient element into the subject position (Flores d'Arcais, 1975).

From their findings, Griffin and Bock proposed that "apprehension precedes formulation", arguing that a holistic process of conceptualization sets the stage for the creation of a to-be-spoken sentence. According to them, the language production process begins with apprehension and generation of a message (a context that language conveys) and proceeds through incremental formulation, with eye movements indicating the temporal relationship between these transitions. More precisely, Griffin and Bock divided the process of speech production upon scene perception into three temporal stages: (1) event apprehension (extracting a coarse understanding of the event as a whole), (2) sentence formulation (the cognitive preparation of linguistic elements, including retrieving and arranging words), and (3) speech execution (overt production). The first stage is proposed to happen fairly quickly, within 300ms from the perception of a scene to describe (Bock, Irwin, Davidson, & Levelt, 2003), which might be attributable to the type of visual stimuli they used – in the cases where scenes

with more complex events are used (as in our experiments), the required time for the first stage would be longer.

This view has been further supported by Levelt and Meyer (2000) (for a similar study, see Meyer, 2004) who showed speakers pictures of objects appearing in different sizes and colors. Speakers named the objects in simple noun phrases such as *the scooter and the ball*, or in a different block of trials, mentioned the size and color of the objects together, as in *the big red scooter and the ball*. Although the speech onset latencies for the two phrase types did not differ significantly (713ms for the simple phrase, 755ms for the complex one), the mean viewing time for the target objects (e.g. scooter) was more than twice as long when complex rather than simple noun phrases were required (559ms for the simple phrase and 1229ms for the complex one). This indicates that when speakers produce complex noun phrases, their eyes remain on the referent object until they have fully planned the phrase to the point of initiating the phrase-final word. Here it seems that planning of sentential structure precedes production of speech.

Nonetheless, the work of Levelt and Meyer indicates that speech production may also be an incremental process. The similar speech onset time for the two phrase types suggests that speakers may initiate uttering complex phrases before having planned all of their constituents. In fact, another line of studies has emphasized the incrementality of language production.

Tomlin (1997) repeatedly showed participants short cartoons of one fish eating another while an arrow pointed to a particular fish, and participants were instructed to keep their eyes on that fish during the presentation. Participants tended to mention the indicated fish first, choosing it as the subject even when they had to use the disfavored passive structure (e.g. "*The red fish is being eaten by the blue fish*"), suggesting that at least in some highly constrained situations, visual attention influences sentence formulation.

Griffin (2001), moreover, conducted an experiment where speakers were required to produce the sentence frame "*The A and the B are above the C*" to describe three pictured objects while object B or C varied in codability (i.e. a measurement of the distribution and frequency of alternative names) and in frequency (i.e. how often the name is used). Although speakers gazed longer at objects with lower codability and lower frequency, their naming onset latency of A was not affected, suggesting that speakers began utterance once a name had been prepared for A, before selecting names for B and C. This implies that a highly incremental process of language production has been used for the task.

More recently, Gleitman and her colleagues (2007) offered a similar account. Contra the earlier mentioned study by Griffin and Bock (2000), they argued that there is no evidence for an initial visual apprehension stage during which the language processing system is disengaged. The authors showed simple pictured events to subjects, which are line drawings of simple transitive and intransitive events similar to the visual stimuli used by Griffin and Bock (2000), and asked them to describe the events while their eye movements were recorded. The subject's initial attention was directed to one character or the other in the events by briefly flashing (60~80ms) a spatial cue just before the onset of each image. The authors examined how manipulations of visual attention affected speakers' linguistic choices when describing scenes, and reported that word order choices appeared to be influenced by early endogenous shifts in attention. They concluded that there is a reliable relationship between initial looking patterns and speaking patterns (i.e. what is attended first is likely to be described first), supporting the incrementality of language production.

However, one should note that the extent to which all of the above studies support incrementality in speech production is somewhat limited. Firstly, Tomlin (1997) used highly restricted experimental settings where subjects were required to focus

on the marked fish during the experiment. The sentence frame and the object arrangement in the display used by Griffin (2001) were fixed throughout the experiment, and this might have encouraged subjects to use incremental strategies. Finally, the findings by Gleitman and her colleagues (2007) demonstrated a correlation between looking patterns and speaking patterns only at the "initial" stage of production, without further assessing whether the relationship extended to the entire production period.

These two positions on the relation between perception and language production – one claims that the holistic conceptual structure mediates the language production process whereas the other claims that the order of perceptual and conceptual input directly influences the linguistic output – have long been debated. We may summarize those two camps as follows:

- ***Structural view*:** Scene comprehension comes before sentence formulation in the way that the sentential structure is determined by the conceptual structure of a scene rather than the perceptual prominence of individual items (e.g. Bock, et al., 2004; Griffin & Bock, 2000; Lashley, 1951).
- ***Incremental view*:** Sentence formulation is done concurrently with scene comprehension so that the sentential structure is built in an incremental manner, while potentially influenced by the perceptual status of each individual item in a scene. (e.g. Gleitman, et al., 2007; Osgood, 1977; Tomlin, 1997).

The two views seemingly describe mutually exclusive principles. However, language production may generally involve both an incremental and a preplanning mechanism (e.g. Levelt, 1989), and the production system may shift between these mechanisms based on the perceived information (Brown-Schmidt & Tanenhaus, 2006).

For instance, Ferreira and Swets (2002) reported experiment results indicating the flexibility of the language system in selecting different policies. In the experiments, they asked subjects to report the sum of digits in three different utterance types: the sum only, "*X is the sum*", and "*the sum is X*". Although subjects can use the third type to spare time for calculating the sum after the utterance onset, the utterance latencies were the same for all three utterance types. Instead, the difficulty of the calculation affected the latencies as well as the duration of the utterance, suggesting that planning precedes speaking in this task. However, they conducted a successive experiment where subjects were instructed to use only the third sentence type while they were pressured to begin to speak quickly – a timing bar was displayed on the screen, and if the sum is not uttered before the timing bar counted all the way down, a loud "beep" sound was produced. In this case, interestingly, both the latencies and the durations were influenced by the difficulty, suggesting that subjects adopted a strategy of simultaneous planning and speaking. Ferreira and Swets concluded that the language production system is not architecturally incremental, but it also at least partly has a capacity to allow planning to occur during articulating.

Moreover, it has been suggested that the language production system is neither solely incremental nor solely structural, based on a study using the Odawa language (Christianson & Ferreira, 2005). Odawa has a rich inventory of verb forms which allows constituents to be freely ordered within the clause, making all logical word orders possible (e.g. VSO, VOS, SVO, OVS, SOV, OSV). In order to examine the production process in Odawa, the authors conducted an experiment in which native Odawa speakers were asked to answer questions that would emphasize different constituents in simple transitive events as the topic (e.g. "*what is happening here?*", "*what is the boy doing?*", "*what is happening to the girl?*"). The experiment results were compatible with a "weaker" version of incrementality; subjects appeared to prefer making highlighted constituents the syntactic subjects by varying verb forms according to the types of questions, but at the same time,

they appeared to avoid simply placing those constituents in the sentence-initial position as shown by their preference for the canonical word order (SVO) even if they had to resort to a less frequent verb form (passive). The authors concluded that the production system strives to encode agents, topics, and humans as subjects while choosing a syntactic frame that will allow for as close as possible to a full alignment of these features to obtain.

Therefore, the stance taken here is that the structural and incremental views do not address two mutually exclusive principles, but rather they are outcomes of two extreme cases. Various perceptual and linguistic factors can drive the language system to switch between or stay in the middle of these two cases. Within the range of the current framework, as we demonstrate in our discussion of TCG in the next section, the relevant factors include a variety of scene properties and task requirements, such as perceptual properties of a scene, scene display time, sentential structure requirements, or time pressure in formulating a sentence. Depending on the combinations of such factors, the system may manifest different styles of behavior which range from a radically incremental style to a strictly structural style – the eye gaze and utterance patterns produced by the system can be controlled by manipulating those perceptual and linguistic factors.

A number of studies indeed demonstrated such variations. Bock and colleagues (2003) reported an effect of perceptual properties on linguistic production by using eye-tracking measures. They showed clock displays in either the number-free analog format or the digital format to subjects and assessed their responses in two categories – the "relative" system includes such expressions as *ten past two* and *quarter (or fifteen) to four*, whereas the "absolute" system includes the corresponding expressions of *two-ten* and *three forty-five*. The results indicated that when subjects produced absolute expressions, they used an incremental strategy (i.e. the speech onset latencies were shorter, and the fixation location and the produced terms showed a tight temporal correlation) whereas when they produced relative expressions, they appeared to prepare the whole expression in advance (i.e. the speech onset latencies were longer with little correlation between the fixation location and the produced terms). It was reported that subjects more easily (less eye-voice span) produced absolute expressions with displays of digital clocks while relative expressions were more easily produced with analog clock displays, thus implying that the perceptual and linguistic compatibility played a role in the choice of the production strategy. In a similar experiment where only analog clock displays were used (Bock, et al., 2004), relative expressions were triggered when the minute hand was in the upper-left quadrant. The result of this experiment also implies the perceptual influence in the selection of the utterance structure and the following incrementability.

A more crucial example is an experiment conducted by van der Meulen (2003). In her experiment, speakers viewed four pictured objects arranged in a square. When the bottom two objects were identical, plural nouns were used ("*A and B are above Cs*") whereas different objects at the bottom called for a two-clause construction ("*A is above B and C is above D*"). When the two types of displays were presented in separate blocks of the experiment (i.e. subjects were sure which sentential structure was to be used), speakers immediately gazed at the first object without scanning the other, resulting in a tight temporal relationship between the gaze and utterance. On the other hand, if the types of displays were intermixed in the same experiment block (i.e. subjects could not be sure which sentential structure was to be used), speakers looked at the bottom objects before starting a name-related gaze on the first object to be mentioned, indicating that speakers employ a flexible strategy in event apprehension and sentence formulation. Thus, the implication is that speakers' perceptual and linguistic policy in scene description is influenced by the task and the settings of the perceived scene.

Recently, moreover, Kuchinsky (2009) claimed that conceptual properties of perceived events matter in the formulation of descriptive utterances at least at the initial stage. She examined the effectiveness of briefly flashing a spatial cue at a character in an event in relation to utterance formulation – i.e. whether or not the cued character is mentioned in the subject position. She divided depicted events into four categories depending on the character and event codability: easy event with easy objects, easy event with difficult objects, difficult event with easy objects, and difficult event with difficult objects. Interestingly, the effect of cuing was found to be significant only for scenes with a difficult event with easy objects. Given that the cuing effect can be interpreted as an incremental strategy in language production (Gleitman, et al., 2007), the implication is that the selection of the strategy makes use of rich perceptual or conceptual information provided by a scene.

5.5. Case Study: Integrating Two Views

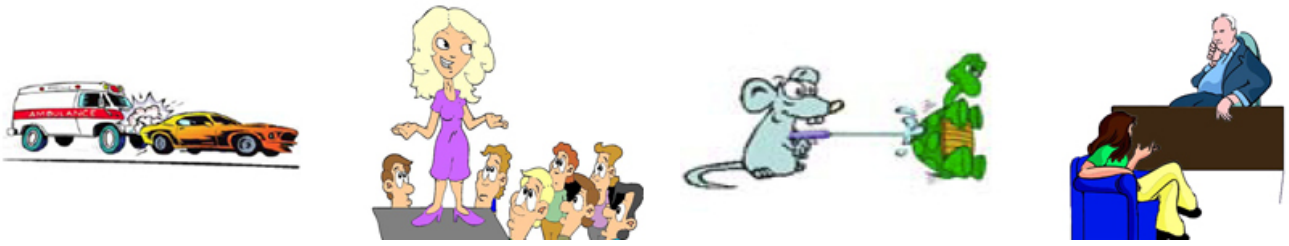
In this section, we discuss detailed accounts on various combinations of perceptual and linguistic factors that drive the production system, especially in regard to the two opposing views addressed in Section 5.4. More specifically, we focus on how eye gaze and utterance patterns corresponding to the incremental and the structural view are generated by the production system of TCG while emphasizing how various experimental constraints and perceptual factors interacting to yield different outcomes.

Among all the relevant factors, we focus on two types of experimental parameters, which are the (conceptual and perceptual) properties of a presented scene and the time pressure given to speakers. This is because we explore distinctive aspects in the process of describing a natural scene by using different settings of eye-tracking experiments in this chapter, and these two parameters are proposed to affect the formation of a SemRep and the production process of TCG (see Section 5.2 and 5.3 for details).

Firstly, it is proposed that the availability of an immediate layout of a perceived scene – i.e. whether or not a certain event (or gist) of the scene is immediately recognizable – determines the process by which a subscene is perceived and encapsulated into a SemRep (Section 2.7). Although it did not directly address the immediate availability of a layout, an analysis of the experiment data suggested that the event complexity and the thematic arrangement of a presented scene affected an initially perceived subscene as reflected by the content of initially made utterances. This implies that the availability of an immediate layout (or gist) can be influenced by the conceptual and perceptual properties of a perceived scene, thus suggesting that the process of building a SemRep is also influenced by these properties.

Moreover, the time pressure given to speakers is proposed to affect their threshold of utterance (Section 4.5), which is one of the key factors of the production process of TCG. Given that threshold is defined as the upper bound of computational resources for generating utterances, the association between threshold and time pressure seems straight forward. The formula of threshold indeed contains a time term, highlighting the linkage between threshold and time pressure. An analysis of the data suggested the close relationship between the well-formedness of an utterance and threshold – low time pressure generally yielded more well-formed utterances whereas high time pressure often resulted in the production of more fragmented (or less well-formed) utterances.

In order to provide detailed accounts on how the system of TCG can address the incremental and structural views, we now analyze an example case of each of view in terms of the perception process of a subscene and the level of threshold.



(A) Easy event with easy characters | **(B) Difficult event with easy characters** | **(C) Easy event with difficult characters** | **(D) Difficult event with difficult characters**

Figure 5.5-1: Example scenes used by Kuchinsky (2009), which are categorized into four types of depicted events depending on codability. The cuing effect was found only for type (B).

Specifically, we examine the study by Kuchinsky (2009) as it previously provided an overall account on the two opposing views, in which she addressed Griffin and Bock (2000) for the structural view and Gleitman et al. (2007) for the incremental view. In her study, Kuchinsky discussed the effects of conceptual and perceptual properties of a perceived event in the early-stage of the production strategy by examining the effect of briefly flashing a spatial cue at a character in a depicted event in utterance formulation (cue flashed 120ms before the display onset). She categorized scenes into four different types depending on the event and character codability (Figure 5.5-1), and found the effect of cuing – i.e. the cued character is mentioned in the subject position – only when the depicted event was difficult to comprehend while event characters were easy to name. During experiments, subjects were asked to "*describe the scene with a single complete sentence as quickly and accurately as possible while avoiding disfluencies*", and only single transitive sentences (e.g. "*the woman is talking to an audience*") were counted into the results while other forms of sentences were discarded.

Kuchinsky's finding indicates that speakers choose an incremental strategy (as evidenced by the cuing effect) in describing a scene when the scene is difficult to comprehend while characters in the scene are easy to recognize, and in other cases, speakers stick to a structural strategy (as indicated by no cuing effect) in which they preplan the sentence structure before starting utterance. From the perspective of the current framework of TCG and SemRep, the case where speakers choose an incremental strategy can be interpreted as follows:

1) The threshold value for the production process during the experiment is assumed to be set relatively low. Considering that the task given to subjects is to produce utterance "as quickly as possible", we may assume that the high time pressure is given during the experiments, and it is generally associated with a low threshold value.
2) The layout of the event is not immediately available since the event is not easy to comprehend. This may lead to building a subscene in an incremental manner by focusing on each characters of the event (subscene extension; Figure 2.7-3). This is supported by a significantly high rate of the intransitive sentence production for difficult event cases (Table 5.5-2).
3) Since characters are easily recognizable, attending to a character immediately results in the recognition of the character. This leads to creating a node in a SemRep, and the successive invocation of a construction instance (possibly a name) attached to the node.

4) A low threshold value drives the system to produce an utterance (possibly a name) once a construction instance is attached. Note that TCG allows an incomplete sentence to be produced (the premature production principle; Section 4.5).
5) The production of the following utterance (possibly a sentential structure from higher-level construction instances) is biased to cope with the already produced utterance in (4), resulting in an incremental process (the utterance continuity principle; Section 4.5).

As shown above, all three conditions (i.e. low threshold, a difficult event, and easy characters) are needed to be met in order for the system to demonstrate a pattern similar to an incremental strategy. Other combinations of the conditional factors (e.g. low threshold, an easy event, and difficult characters) are not guaranteed to yield an incremental pattern, but rather may result in a structural pattern.

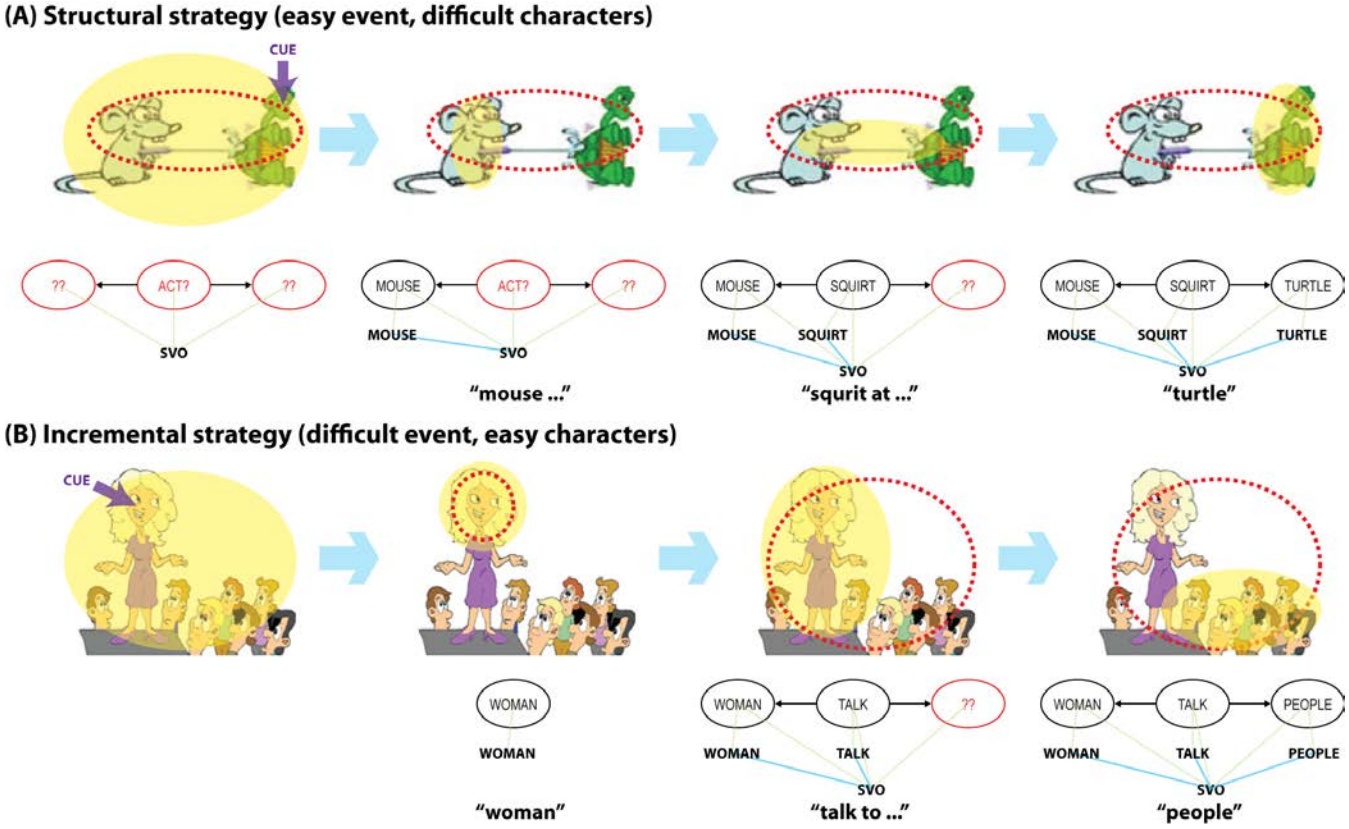


Figure 5.5-2: An illustration of "hypothesized" TCG processes in two cases of Kuchinsky (2009)'s study (see Appendices for the simulation output). (A) illustrates the case where the sentence structure is specified first, resulting in a production process consistent with the structural view (i.e. no cueing effect due to the verbal guidance principle; Section 4.5), whereas (B) illustrates the case where a lexical item ("*woman*") is specified earlier than the sentential structure, resulting in a process consistent with the incremental view (i.e. cueing effect). Since low threshold is assumed for both cases, utterances are being made intermittently.

Table 5.5-1 summarizes the behaviors of the system with all possible combinations of the relevant factors. The perceptual difficulty column represents the relative difficulty between an event and characters – e.g. difficult event with

difficult characters is represented as "Event = Characters", and easy event with difficult characters is represented as "Event < Characters", etc. Note that the outcome pattern of the system is specified as "structural" when the difficulty of an event and characters are the same. This is based on a general tendency of speakers towards a structural strategy (especially when they have a sentential structure available at the moment) – *people don't start what they can't finish* (Bock, et al., 2004). All high threshold cases yield a structural pattern due to the same reason.

Table 5.5-1: All combinations of the relevant factors in the process of scene perception and description in TCG.

Threshold	Perceptual Difficulty	Subscene Perception	Available Construction	Outcome Pattern
Low	Event < Characters	Specification	Sentential structures first	Structural
	Event > Characters	Extension	Lexical items first	Incremental
	Event = Characters	Depending on salience	Lexical items and sentential structures	Structural
High	Event < Characters	Specification	Sentential structures first	Structural
	Event > Characters	Extension	Lexical items first	Structural
	Event = Characters	Depending on salience	Lexical items and sentential structures	Structural

Generally speaking, the difficulty of an event relative to the difficulty of its characters decides the "order" in which the lexical items and the sentential structures are ready – e.g. an easy event coupled with difficult characters may result in the sentential structure being ready first, whereas for a difficult event coupled with easy characters, the lexical items may be ready first. Moreover, threshold decides the "time" to produce an utterance – e.g. if threshold is low, the system produces fragmented lexical items even if the sentential structure is not ready yet, whereas if threshold is high the system does not produce an utterance even if lexical items are ready first, allowing the later-specified high-level constructions to decide the sentential structure.

Table 5.5-2: The percentage of response types split by event and character accessibility (adapted from Table 7 of Kuchinsky, 2009).

Event Accessibility	First-Term Accessibility	Response Type		
		Two-Event	Intransitive	Single Transitive Event (Scored)
High (Easy)	High (Easy)	4.11%	8.90%	86.99%
	Low (Hard)	3.49%	16.86%	79.65%
Low (Hard)	High (Easy)	7.78%	34.44%	57.78%
	Low (Hard)	10.34%	52.79%	36.87%

Transitive (scored): "*The woman is speaking to an audience.*"

Two-event: "*The woman is speaking and men are watching.*"

Intransitive: "*The woman is speaking.*"

Moreover, Kuchinsky reported a significantly high rate of the intransitive sentence production for difficult event cases than easy event cases (Table 5.5-2), even though subjects were instructed to refrain from producing such sentences. Similar to what is indicated in the analysis of the experiment data presented in Section 5.3, this data can also be interpreted in terms of the influence of the scene properties on the perception process of a (initial) subscene since the intransitive sentence (e.g. "*The woman is speaking*") involves less detail compared to the transitive sentence (e.g. "*The woman is speaking to an audience*") or the two-event sentences (e.g. "*The woman is speaking and men are watching*"), implying the smaller conceptual and perceptual coverage of subscenes for more difficult events.

(A)  **(B)** 

Figure 5.5-3: Example stimulus scenes used by Griffin and Bock (2000) (A) and Gleitman et al. (2007) (B).

The studies of Griffin and Bock (2000) and Gleitman et al. (2007) differed most in their scene properties. Griffin and Bock mostly used scenes with easily recognizable events, which are proposed to be associated with a structural strategy, whereas Gleitman et al. used scenes with somewhat ambiguous settings, proposed to be associated with an incremental strategy (Figure 5.5-1). In particular, the event structure of most of Gleitman et al.'s stimuli (even transitive events) does not appear to strictly constrain the voice of the sentence. Although it cannot be said exactly how much such a difference mattered in the analysis of each of these studies (e.g. event/character properties were not analyzed, and there was no time pressure to subjects in both of the studies, etc.), subjects' production strategy might have been affected to some extent.

So far, we have addressed studies with specific examples in an account for the structural view and the incremental view. Although these studies well address each of these views in an early stage of the process, the coverage of these studies are fairly limited when it comes to the entire duration of the scene description process. In order to fully address the implications of the specific strategies that speakers take, the current study needs to be extended beyond the "starting point" to include the eye gaze and utterance patterns covering the entire process of scene description.

Chapter 6. Conclusion

Throughout the thesis, we have explored the possible mechanisms of the vision system, through which a SemRep is encapsulated from a perceived scene, and addressed the theoretical framework and the implementation details of TCG, which we propose as the language model that exploits the structural characteristics of the particular representation format of SemRep. We have also related TCG to the experimental analysis of how descriptions of a scene may be influenced under different time constraints and perceptual properties.

Although the current work provides a detailed account on the dynamic interplay between the visual perception and speech production, we envision several interesting extensions for future work.

Firstly, the implemented model runs on a relative time frame, without making any strong argument on actual temporal transitions involved in processing perceived visual information and applying constructions. The current model avoided the question by simply assuming that the simulation time used in the implemented system is generally assumed to correspond to a "cognitively important" transition of the conceptual status of a speaker. In future work, however, specific details on the time scales of the procedures, through which the perceived visual information (i.e. scene) is translated into a verbal expression, will be more fully discussed – a number of recent studies suggested fine-grained details in speech planning and its scope (Allum & Wheeldon, 2007; R. C. Martin, et al., 2010; Oppermann, et al., 2010). Moreover, as proposed by other studies (Andersson, et al., 2006; Griffin & Bock, 2000; van der Meulen, et al., 2001), we also found a general pattern of "apprehension-production cycle" from the experimental data, in which speakers tend to look at the object just before naming the object. However, this has not been discussed in the current work. Therefore, the future extension of TCG will address this issue more fully and provide specific computational mechanisms that can work in the real-time domain.

Moreover, the simulation results presented in Section 4.6 are still "preliminary" in the sense that the analysis results in the temporal dispositions of subjects' eye movements are not reflected in designing the scene description files used for simulations. This is partly due to the fact that the current implementation model uses a relative time frame, which is the shortcoming of the current model addressed earlier. In future work, we will perform detailed measurements on the temporal transitions in subjects' performance during the task of scene description and find a general pattern of subjects' fixations in relation to the produced utterances (e.g. apprehension-production cycle). By doing so, we will be able to provide both a good general argument and a more compelling set of simulation output that shows how different conditions can be seen to explain the data in a way consistent with the currently proposed assessment on how to bridge the divide between the structural and incremental view.

Another aspect of the current version of TCG that can be a candidate for future extension is its computational framework. Motivated by the VISIONS system and schema theory, TCG is currently built as a system in which competition and cooperation between schema instances (i.e. constructions) generate a verbal description of a perceived visual scene. The HEARSAY speech understanding system (Erman, Hayes-Roth, Lesser, & Reddy, 1980) provides a very similar 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. 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. This neuro-HEARSAY may provide one inspiration for the direction, via which future work should move beyond the current work in developing neural models of the interaction of SemRep and TCG in the process of producing an utterance of the given semantics (SemRep), or vice versa.

Lastly, 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 as well. During production of utterances, a given graph is compared with a number of constructions for similarity. Only the winner is to be chosen to produce utterances. On the other hand, in comprehension mode, a textual form, which is considered to be "heard" by the system, activates constructions by an inverse matching mechanism. When proper constructions are chosen by matching the Syn-Froms of constructions, a new SemRep graph, which is regarded as the end-production of the comprehension process, would be built from the Sem-Frames of the constructions. In this case, the same set of constructions can be used as well. However, this requires an extension on a computation paradigm for semantic processing (in terms of SemRep) to go beyond the linguistic domain. Such extension is basically 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. Moreover, analysis results on agrammatic aphasics also revealed that they could still correctly process aspectual and complement coercion operations, thought to be purely semantic in nature, while ignoring syntactic cues (Piñango & Zurif, 2001). Indeed, Piñango (2006) stressed that comprehension can take place despite syntactic impairment, but only if the sentence's semantic structure is rich enough.

Based on this account, we recently provided the conceptual framework of the comprehension model of TCG (Barrès & Lee, 2013). We introduced a theoretical distinction between the world knowledge "heavy semantics" that survives in agrammatic aphasics and the "light semantics" of syntactico-semantic categories that corresponds to the semantics of slot-fillers in grammatical constructions in Construction Grammar. The model is described as a two-route system where the light semantics path interprets an utterance through a parsing process which yields a pyramid of constructions from which a SemRep may be read off, whereas the heavy semantics path exploits the same processes that can build a SemRep during perception and action planning. A process of competition and cooperation couples the two paths to constrain heavy semantics by grammatical cues so that a SemRep which has a few nodes created by recognition of content words can be expanded to add missing nodes and edges suggested by world knowledge.

Appendices

Appendix A. Semantic Network

The following is the full definition of the semantic network used for the simulations in the current work.

```
#
# TCG Semantic Network definition
#


############################################################
# relation-related semantics
############################################################

is_a RELATION
{
            MODIFY
            THEMATIC
            {
                        AGENT
                        PATIENT
            }
            TEMPORAL
            {
                        SUCCESSIVE
                        CONCURRENT
            }
            SPATIAL { IN }
}


############################################################
# object-related semantics
############################################################

is_a ENTITY
{
            OBJECT
            {
                        HUMAN
                        {
                                    MAN BOY
                                    WOMAN GIRL
                                    PEOPLE
                        }
                        ITEM
                        {
                                    CLOTHING
                                    {
                                                DRESS
                                                TSHIRT
                                    }
                        }
                        ANIMAL
                        {
                                    MOUSE TURTLE
                        }
            }
            PLACE
            {
                        BOXINGRING
                        PARK
            }
}

is_a ANIMATE { HUMAN ANIMAL }

is_a MALE { MAN BOY }
is_a FEMALE { WOMAN GIRL }


############################################################
# action-related semantics
############################################################

is_a ACTION
{
            TRANSITIVE
            {
                        SQUIRT
                        TALK
                        HIT
                        KICK
                        WEAR
            }
            INTRANSITIVE
            {
                        LAUGH
                        WATCH
            }
}


############################################################
# property-related semantics
############################################################
```

```
is_a PROPERTY
{
            COLOR
            {
                        BLACK
                        BLUE
                        GREEN
            }
            SIZE
            {
                        SMALL
                        BIG
            }
            APPEARANCE
            {
                        PRETTY
                        HANDSOME
            }
}
```

Appendix B. Construction Set

The following is the entire construction definitions used for the simulations in the current work.

```
#
# TCG Construction Vocabulary
#


############################################################
# sentence structures
############################################################

construction CNJ_AND
{
            class: S

            node EVT1 { concept: ACTION+ shared head }
            node EVT2 { concept: ACTION+ shared head }
            relation EVT1_EVT2 { concept: SUCCESSIVE from: EVT1 to: EVT2 }

            [EVT1: S] 'and' [EVT2: S]
}

construction CNJ_WHILE
{
            class: S

            node EVT1 { concept: ACTION+ shared head }
            node EVT2 { concept: ACTION+ shared head }
            relation EVT1_EVT2 { concept: CONCURRENT from: EVT1 to: EVT2 }

            [EVT1: S] 'while' [EVT2: S]
}

construction SVO
{
            class: S
            preference: 1           # sentential structure preference

            node SUBJ { concept: ENTITY+ shared head }
            node OBJ { concept: ENTITY+ shared }
            node ACT { concept: ACTION+ shared head }
            relation ACT_SUBJ { concept: AGENT from: ACT to: SUBJ }
            relation ACT_OBJ { concept: PATIENT from: ACT to: OBJ }

            [SUBJ: NC NP N] [ACT: VP V] [OBJ: NC NP N]
}

construction PAS_SVO
{
            class: S

            node SUBJ { concept: ENTITY+ shared }
            node OBJ { concept: ENTITY+ shared head }
            node ACT { concept: ACTION+ shared head }
            relation ACT_SUBJ { concept: AGENT from: ACT to: SUBJ }
            relation ACT_OBJ { concept: PATIENT from: ACT to: OBJ }

            [OBJ: NC NP N] 'is' [ACT: VP V] '-ed by' [SUBJ: NC NP N]
}

construction EXIST_S
{
            class: S
            preference: 1           # sentential structure preference

            node SUBJ { concept: OBJECT+ shared head }

            'there is' [SUBJ: NC NP N]
}

construction THEME_S
{
            class: S
            preference: 1           # sentential structure preference

            node SUBJ { concept: PLACE+ shared head }
```

```
            'it is' [SUBJ: NC NP N]
}

construction SPA
{
            class: S
            preference: 1           # sentential structure preference

            node OBJ { concept: ENTITY+ shared head }
            node ATTR { concept: PROPERTY+ shared }
            relation ATTR_OBJ { concept: MODIFY from: ATTR to: OBJ }

            [OBJ: NC NP N] 'is' [ATTR: A]
}

construction SV
{
            class: S
            preference: 1           # sentential structure preference

            node SUBJ { concept: ENTITY+ shared head }
            node ACT { concept: INTRANSITIVE+ shared head }
            relation ACT_SUBJ { concept: AGENT from: ACT to: SUBJ }

            [SUBJ: NC NP N] [ACT: VP V]
}

construction PP_IN
{
            class: S
            preference: 1           # sentential structure preference

            node EVT { concept: ACTION+ shared head }
            relation EVT_PP { concept: IN from: EVT to: PLACE }
            node PLACE { concept: ENTITY+ shared }

            [EVT: S] 'in' [PLACE: NC NP N]
}



############################################################
# complex constructions
############################################################

construction REL_SVO_WHO
{
            class: NC

            node SUBJ { concept: HUMAN+ shared head }
            node OBJ { concept: ENTITY+ shared }
            node ACTION { concept: ACTION+ shared }
            relation ACTION_SUBJ { concept: AGENT from: ACTION to: SUBJ }
            relation ACTION_OBJ { concept: PATIENT from: ACTION to: OBJ }

            [SUBJ: NP N] 'who' [ACTION: VP V] [OBJ: NC NP N]
}

construction REL_SVO_WHICH
{
            class: NC

            node SUBJ { concept: ITEM+ shared head }
            node OBJ { concept: ENTITY+ shared }
            node ACTION { concept: ACTION+ shared }
            relation ACTION_SUBJ { concept: AGENT from: ACTION to: SUBJ }
            relation ACTION_OBJ { concept: PATIENT from: ACTION to: OBJ }

            [SUBJ: NP N] 'which' [ACTION: VP V] [OBJ: NC NP N]
}

construction REL_SV_WHO
{
            class: NC

            node SUBJ { concept: HUMAN+ shared head }
            node ACTION { concept: INTRANSITIVE+ shared }
            relation ACTION_SUBJ { concept: AGENT from: ACTION to: SUBJ }

            [SUBJ: NP N] 'who' [ACTION: VP V]
}

construction REL_SV_WHICH
{
            class: NC

            node SUBJ { concept: ITEM+ shared head }
            node ACTION { concept: INTRANSITIVE+ shared }
            relation ACTION_SUBJ { concept: AGENT from: ACTION to: SUBJ }

            [SUBJ: NP N] 'which' [ACTION: VP V]
}

construction REL_PAS_SVO_WHO
{
            class: NC

            node SUBJ { concept: ENTITY+ shared }
            node OBJ { concept: HUMAN+ shared head }
            node ACTION { concept: ACTION+ shared }
            relation ACTION_SUBJ { concept: AGENT from: ACTION to: SUBJ }
            relation ACTION_OBJ { concept: PATIENT from: ACTION to: OBJ }

            [OBJ: NP N] 'who is' [ACTION: VP V] '-ed by' [SUBJ: NC NP N]
}
```

```
construction REL_SPA_WHO
{
            class: NC

            node OBJ { concept: HUMAN+ shared head }
            node ATTR { concept: PROPERTY+ shared }
            relation ATTR_OBJ { concept: MODIFY from: ATTR to: OBJ }

            [OBJ: NP N] 'who is' [ATTR: A]
}

construction REL_SPA_WHICH
{
            class: NC

            node OBJ { concept: ITEM+ shared head }
            node ATTR { concept: PROPERTY+ shared }
            relation ATTR_OBJ { concept: MODIFY from: ATTR to: OBJ }

            [OBJ: NP N] 'which is' [ATTR: A]
}

construction ADJ_NOUN
{
            class: NP

            node OBJ { concept: ENTITY+ shared head }
            node ATTR { concept: PROPERTY+ shared }
            relation ATTR_OBJ { concept: MODIFY from: ATTR to: OBJ }

            [ATTR: A] [OBJ: NP N]
}

construction IN_COLOR
{
            class: NP

            node HUMAN { concept: HUMAN+ shared head }
            node WEAR { concept: WEAR }
            node CLOTH { concept: CLOTHING+ }
            node COLOR { concept: COLOR+ shared }
            relation HUMAN_WEAR { concept: AGENT from: WEAR to: HUMAN }
            relation CLOTH_WEAR { concept: PATIENT from: WEAR to: CLOTH }
            relation COLOR_CLOTH { concept: MODIFY from: COLOR to: CLOTH }

            [HUMAN: NP N] 'in' [COLOR: A]
}

############################################################
# lexicons
############################################################

# verbs
construction HIT         { class: V node NODE { concept: HIT head } 'hit' }
construction KICK        { class: V node NODE { concept: KICK head } 'kick' }
construction WEAR        { class: V node NODE { concept: WEAR head } 'wear' }
construction LAUGH       { class: V node NODE { concept: LAUGH head } 'laugh' }
construction WATCH       { class: V node NODE { concept: WATCH head } 'watch' }
construction TALK        { class: V node NODE { concept: TALK head } 'talk to' }
construction SQUIRT      { class: V node NODE { concept: SQUIRT head } 'squirt at' }

# adjectives
construction BLUE        { class: A node NODE { concept: BLUE head } 'blue' }
construction BLACK       { class: A node NODE { concept: BLACK head } 'black' }
construction GREEN       { class: A node NODE { concept: GREEN head } 'green' }

construction SMALL       { class: A node NODE { concept: SMALL head } 'small' }
construction BIG         { class: A node NODE { concept: BIG head } 'big' }

construction HANDSOME    { class: A node NODE { concept: HANDSOME head } 'handsome' }
construction PRETTY      { class: A node NODE { concept: PRETTY head } 'pretty' }

# nouns
construction WOMAN       { class: N node NODE { concept: WOMAN head } 'woman' }
construction MAN         { class: N node NODE { concept: MAN head } 'man' }
construction GIRL        { class: N node NODE { concept: GIRL head } 'girl' }
construction BOY         { class: N node NODE { concept: BOY head } 'boy' }
construction PEOPLE      { class: N node NODE { concept: PEOPLE head } 'people' }

construction MOUSE       { class: N node NODE { concept: MOUSE head } 'mouse' }
construction TURTLE                  { class: N node NODE { concept: TURTLE head } 'turtle' }

construction DRESS       { class: N node NODE { concept: DRESS head } 'dress' }
construction TSHIRT      { class: N node NODE { concept: TSHIRT head } 't-shirt' }

construction BOXINGRING { class: N node NODE { concept: BOXINGRING head } 'boxing ring' }
construction PARK        { class: N node NODE { concept: PARK head } 'park' }
```

Appendix C. High and Low Threshold Cases

In Section 4.5, two sets of computational stages of the TCG process for a high and low threshold case are illustrated. This appendix provides the simulation results corresponding to each of those cases.

The following is the scene description file used for the simulation, in which the perceptual schemas for the semantics WOMAN, WOMAN-HIT-MAN, PRETTY-WOMAN, WOMAN-WEAR-DRESS, and BLUE-DRESS are successively

perceived for updating the SemRep.

```
#
# TCG Scene: Woman-hit-man
#
# the famous "pretty woman in blue hit man" example
#

image: "woman hits man.jpg"
resolution: 400 * 326

region WOMAN_AREA
{
            location: 283, 205 size: 50, 150
            saliency: 100
            uncertainty: 1

            object WOMAN { concept: WOMAN }

            perceive WOMAN
}

region HIT_AREA
{
            location: 213, 110 size: 60, 20
            saliency: 90
            uncertainty: 1

            object MAN { concept: MAN }
            object HIT { concept: HIT }
            relation HIT_AGENT { concept: AGENT from: HIT to: WOMAN }
            relation HIT_PATIENT { concept: PATIENT from: HIT to: MAN }

            perceive HIT, MAN, HIT_AGENT, HIT_PATIENT
}

region WOMAN_FACE_AREA
{
            location: 283, 205 size: 50, 150
            saliency: 70
            uncertainty: 1

            object PRETTY { concept: PRETTY }
            relation PRETTY_MODIFY { concept: MODIFY from: PRETTY to: WOMAN }

            perceive PRETTY, PRETTY_MODIFY
}

region DRESS_AREA
{
            location: 283, 205 size: 50, 150
            saliency: 50
            uncertainty: 1

            object DRESS { concept: DRESS }
            object WEAR { concept: WEAR }
            relation WEAR_AGENT { concept: AGENT from: WEAR to: WOMAN }
            relation WEAR_PATIENT { concept: PATIENT from: WEAR to: DRESS }

            perceive WEAR, DRESS, WEAR_AGENT, WEAR_PATIENT
}

region DRESS_FOCUS_AREA
{
            location: 283, 205 size: 50, 150
            saliency: 40
            uncertainty: 1

            object BLUE { concept: BLUE }
            relation BLUE_MODIFY { concept: MODIFY from: BLUE to: DRESS }

            perceive BLUE, BLUE_MODIFY
}
```

The following is the simulation output for the high threshold case. In this case, all of the threshold parameters are set to infinite.

```
Template Construction Grammar (TCG) Simulator v2.5

Jinyong Lee (jinyongl@usc.edu), June 23. 2012
USC Brain Project, Computer Science Department
University of Southern California (USC)

Loading Initialization File 'TCG.ini'...
Loading Semantic Network 'TCG_semantics.txt'...
Loading Construction Vocabulary 'TCG_vocabulary.txt'...
Loading Scene 'scene_womanhitman.txt'...

Initializing Simulator...
- Max Simulation Time: 20
- Premature Production: on
- Utterance Continuity: on
- Verbal Guidance: on
- Threshold of Utterance: Time = infinite, CNXs = infinite, Syllables = infinite

Beginning Simulation...

================================================================================
  Simulation Time: 1
================================================================================
> Current Attention
```

```
  None

> Next Attention
  WOMAN_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 2
================================================================================
> Current Attention
  WOMAN_AREA (perception done)

> Perceived Regions
  WOMAN_AREA

> Schema Instances
[!O] SemRep-N WOMAN_0
[!O] Construction EXIST_S_1 covering WOMAN_0 for 'there is' [WOMAN_2]
[!O] Construction WOMAN_2 covering WOMAN_0 for 'woman'

> Construction Structures
[ ] 138: EXIST_S_1 'there is' [WOMAN_2 'woman']

> Next Attention
  HIT_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 3
================================================================================
> Current Attention
  HIT_AREA (perception done)

> Perceived Regions
  HIT_AREA

> Schema Instances
[ O] SemRep-N WOMAN_0
[ O] Construction EXIST_S_1 covering WOMAN_0 for 'there is' [REL_SVO_WHO_10]
[ O] Construction WOMAN_2 covering WOMAN_0 for 'woman'
[!O] SemRep-N HIT_3
[!O] SemRep-N MAN_4
[!O] SemRep-R AGENT_5 from HIT_3 to WOMAN_0
[!O] SemRep-R PATIENT_6 from HIT_3 to MAN_4
[!O] Construction SVO_7 covering WOMAN_0 MAN_4 HIT_3 AGENT_5 PATIENT_6 for [WOMAN_2] [HIT_12] [MAN_13]
[!X] Construction PAS_SVO_8 covering WOMAN_0 MAN_4 HIT_3 AGENT_5 PATIENT_6 for [MAN_13] 'is' [HIT_12] '-ed by' [WOMAN_2]
[!O] Construction EXIST_S_9 covering MAN_4 for 'there is' [REL_PAS_SVO_WHO_11]
[!X] Construction REL_SVO_WHO_10 covering WOMAN_0 MAN_4 HIT_3 AGENT_5 PATIENT_6 for [WOMAN_2] 'who' [HIT_12] [MAN_13]
[!X] Construction REL_PAS_SVO_WHO_11 covering WOMAN_0 MAN_4 HIT_3 AGENT_5 PATIENT_6 for [MAN_13] 'who is' [HIT_12] '-ed by' [WOMAN_2]
[!O] Construction HIT_12 covering HIT_3 for 'hit'
[!O] Construction MAN_13 covering MAN_4 for 'man'

> Competition Traces
  SVO_7(539) eliminated PAS_SVO_8(483)
  SVO_7(539) eliminated REL_SVO_WHO_10(529)
  SVO_7(539) eliminated REL_PAS_SVO_WHO_11(523)

> Construction Structures
[ ] 138: EXIST_S_1 'there is' [WOMAN_2 'woman']
[ ] 140: EXIST_S_9 'there is' [MAN_13 'man']
[ ] 539: SVO_7 [WOMAN_2 'woman'] [HIT_12 'hit'] [MAN_13 'man']
[X] 483: PAS_SVO_8 [MAN_13 'man'] 'is' [HIT_12 'hit'] '-ed by' [WOMAN_2 'woman']
[X] 529: EXIST_S_1 'there is' [REL_SVO_WHO_10 [WOMAN_2 'woman'] 'who' [HIT_12 'hit'] [MAN_13 'man']]
[X] 523: EXIST_S_9 'there is' [REL_PAS_SVO_WHO_11 [MAN_13 'man'] 'who is' [HIT_12 'hit'] '-ed by' [WOMAN_2 'woman']]

> Next Attention
  WOMAN_FACE_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 4
================================================================================
> Current Attention
  WOMAN_FACE_AREA (perception done)

> Perceived Regions
  WOMAN_FACE_AREA

> Schema Instances
[ O] SemRep-N WOMAN_0
[ O] Construction EXIST_S_1 covering WOMAN_0 for 'there is' [ADJ_NOUN_18]
[ O] Construction WOMAN_2 covering WOMAN_0 for 'woman'
[ O] SemRep-N HIT_3
[ O] SemRep-N MAN_4
[ O] SemRep-R AGENT_5 from HIT_3 to WOMAN_0
[ O] SemRep-R PATIENT_6 from HIT_3 to MAN_4
[ O] Construction SVO_7 covering WOMAN_0 MAN_4 HIT_3 AGENT_5 PATIENT_6 for [ADJ_NOUN_18] [HIT_12] [MAN_13]
[ O] Construction EXIST_S_9 covering MAN_4 for 'there is' [MAN_13]
[ O] Construction HIT_12 covering HIT_3 for 'hit'
[ O] Construction MAN_13 covering MAN_4 for 'man'
[!O] SemRep-N PRETTY_14
[!O] SemRep-R MODIFY_15 from PRETTY_14 to WOMAN_0
[!X] Construction SPA_16 covering WOMAN_0 PRETTY_14 MODIFY_15 for [WOMAN_2] 'is' [PRETTY_19]
[!X] Construction REL_SPA_WHO_17 covering WOMAN_0 PRETTY_14 MODIFY_15 for [WOMAN_2] 'who is' [PRETTY_19]
[!O] Construction ADJ_NOUN_18 covering WOMAN_0 PRETTY_14 MODIFY_15 for [PRETTY_19] [WOMAN_2]
[!O] Construction PRETTY_19 covering PRETTY_14 for 'pretty'

> Competition Traces
  REL_SPA_WHO_17(728) eliminated SPA_16(337)
  ADJ_NOUN_18(733) eliminated REL_SPA_WHO_17(728)

> Construction Structures
[ ] 138: EXIST_S_1 'there is' [WOMAN_2 'woman']
[ ] 140: EXIST_S_9 'there is' [MAN_13 'man']
[X] 337: SPA_16 [WOMAN_2 'woman'] 'is' [PRETTY_19 'pretty']
[X] 327: EXIST_S_1 'there is' [REL_SPA_WHO_17 [WOMAN_2 'woman'] 'who is' [PRETTY_19 'pretty']]
[ ] 332: EXIST_S_1 'there is' [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']]
[ ] 539: SVO_7 [WOMAN_2 'woman'] [HIT_12 'hit'] [MAN_13 'man']
[X] 728: SVO_7 [REL_SPA_WHO_17 [WOMAN_2 'woman'] 'who is' [PRETTY_19 'pretty']] [HIT_12 'hit'] [MAN_13 'man']
[ ] 733: SVO_7 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] [HIT_12 'hit'] [MAN_13 'man']
```

```
> Next Attention
  DRESS_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 5
================================================================================
> Current Attention
  DRESS_AREA (perception done)

> Perceived Regions
  DRESS_AREA

> Schema Instances
[ O] SemRep-N WOMAN_0
[ O] Construction EXIST_S_1 covering WOMAN_0 for 'there is' [REL_SVO_WHO_27]
[ O] Construction WOMAN_2 covering WOMAN_0 for 'woman'
[ O] SemRep-N HIT_3
[ O] SemRep-N MAN_4
[ O] SemRep-R AGENT_5 from HIT_3 to WOMAN_0
[ O] SemRep-R PATIENT_6 from HIT_3 to MAN_4
[ O] Construction SVO_7 covering WOMAN_0 MAN_4 HIT_3 AGENT_5 PATIENT_6 for [REL_SVO_WHO_27] [HIT_12] [MAN_13]
[ O] Construction EXIST_S_9 covering MAN_4 for 'there is' [MAN_13]
[ O] Construction HIT_12 covering HIT_3 for 'hit'
[ O] Construction MAN_13 covering MAN_4 for 'man'
[ O] SemRep-N PRETTY_14
[ O] SemRep-R MODIFY_15 from PRETTY_14 to WOMAN_0
[ O] Construction ADJ_NOUN_18 covering WOMAN_0 PRETTY_14 MODIFY_15 for [PRETTY_19] [WOMAN_2]
[ O] Construction PRETTY_19 covering PRETTY_14 for 'pretty'
[!O] SemRep-N WEAR_20
[!O] SemRep-N DRESS_21
[!O] SemRep-R AGENT_22 from WEAR_20 to WOMAN_0
[!O] SemRep-R PATIENT_23 from WEAR_20 to DRESS_21
[!X] Construction SVO_24 covering WOMAN_0 DRESS_21 WEAR_20 AGENT_22 PATIENT_23 for [ADJ_NOUN_18] [WEAR_28] [DRESS_29]
[!X] Construction PAS_SVO_25 covering WOMAN_0 DRESS_21 WEAR_20 AGENT_22 PATIENT_23 for [DRESS_29] 'is' [WEAR_28] '-ed by' [ADJ_NOUN_18]
[!O] Construction EXIST_S_26 covering DRESS_21 for 'there is' [DRESS_29]
[!O] Construction REL_SVO_WHO_27 covering WOMAN_0 DRESS_21 WEAR_20 AGENT_22 PATIENT_23 for [ADJ_NOUN_18] 'who' [WEAR_28] [DRESS_29]
[!O] Construction WEAR_28 covering WEAR_20 for 'wear'
[!O] Construction DRESS_29 covering DRESS_21 for 'dress'

> Competition Traces
  SVO_24(730) eliminated PAS_SVO_25(674)
  REL_SVO_WHO_27(1121) eliminated SVO_24(730)

> Construction Structures
[ ] 138: EXIST_S_1 'there is' [WOMAN_2 'woman']
[ ] 140: EXIST_S_9 'there is' [MAN_13 'man']
[ ] 138: EXIST_S_26 'there is' [DRESS_29 'dress']
[ ] 332: EXIST_S_1 'there is' [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']]
[ ] 539: SVO_7 [WOMAN_2 'woman'] [HIT_12 'hit'] [MAN_13 'man']
[X] 536: SVO_24 [WOMAN_2 'woman'] [WEAR_28 'wear'] [DRESS_29 'dress']
[X] 480: PAS_SVO_25 [DRESS_29 'dress'] 'is' [WEAR_28 'wear'] '-ed by' [WOMAN_2 'woman']
[ ] 733: SVO_7 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] [HIT_12 'hit'] [MAN_13 'man']
[X] 730: SVO_24 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] [WEAR_28 'wear'] [DRESS_29 'dress']
[X] 674: PAS_SVO_25 [DRESS_29 'dress'] 'is' [WEAR_28 'wear'] '-ed by' [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']]
[ ] 526: EXIST_S_1 'there is' [REL_SVO_WHO_27 [WOMAN_2 'woman'] 'who' [WEAR_28 'wear'] [DRESS_29 'dress']]
[ ] 720: EXIST_S_1 'there is' [REL_SVO_WHO_27 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] 'who' [WEAR_28 'wear'] [DRESS_29 'dress']]
[ ] 927: SVO_7 [REL_SVO_WHO_27 [WOMAN_2 'woman'] 'who' [WEAR_28 'wear'] [DRESS_29 'dress']] [HIT_12 'hit'] [MAN_13 'man']
[ ] 1121: SVO_7 [REL_SVO_WHO_27 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] 'who' [WEAR_28 'wear'] [DRESS_29 'dress']] [HIT_12 'hit'] [MAN_13
'man']

> Next Attention
  DRESS_FOCUS_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 6
================================================================================
> Current Attention
  DRESS_FOCUS_AREA (perception done)

> Perceived Regions
  DRESS_FOCUS_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ O] Construction EXIST_S_1 covering WOMAN_0 for 'there is' [IN_COLOR_35]
[ @] Construction WOMAN_2 covering WOMAN_0 for 'woman'
[ @] SemRep-N HIT_3
[ @] SemRep-N MAN_4
[ @] SemRep-R AGENT_5 from HIT_3 to WOMAN_0
[ @] SemRep-R PATIENT_6 from HIT_3 to MAN_4
[ @] Construction SVO_7 covering WOMAN_0 MAN_4 HIT_3 AGENT_5 PATIENT_6 for [IN_COLOR_35] [HIT_12] [MAN_13]
[ O] Construction EXIST_S_9 covering MAN_4 for 'there is' [MAN_13]
[ @] Construction HIT_12 covering HIT_3 for 'hit'
[ @] Construction MAN_13 covering MAN_4 for 'man'
[ @] SemRep-N PRETTY_14
[ @] SemRep-R MODIFY_15 from PRETTY_14 to WOMAN_0
[ @] Construction ADJ_NOUN_18 covering WOMAN_0 PRETTY_14 MODIFY_15 for [PRETTY_19] [WOMAN_2]
[ @] Construction PRETTY_19 covering PRETTY_14 for 'pretty'
[ @] SemRep-N WEAR_20
[ @] SemRep-N DRESS_21
[ @] SemRep-R AGENT_22 from WEAR_20 to WOMAN_0
[ @] SemRep-R PATIENT_23 from WEAR_20 to DRESS_21
[ O] Construction EXIST_S_26 covering DRESS_21 for 'there is' [ADJ_NOUN_34]
[ X] Construction REL_SVO_WHO_27 covering WOMAN_0 DRESS_21 WEAR_20 AGENT_22 PATIENT_23 for [ADJ_NOUN_18] 'who' [WEAR_28] [ADJ_NOUN_34]
[ X] Construction WEAR_28 covering WEAR_20 for 'wear'
[ X] Construction DRESS_29 covering DRESS_21 for 'dress'
[!@] SemRep-N BLUE_30
[!@] SemRep-R MODIFY_31 from BLUE_30 to DRESS_21
[!X] Construction SPA_32 covering DRESS_21 BLUE_30 MODIFY_31 for [DRESS_29] 'is' [BLUE_36]
[!X] Construction REL_SPA_WHICH_33 covering DRESS_21 BLUE_30 MODIFY_31 for [DRESS_29] 'which is' [BLUE_36]
[!X] Construction ADJ_NOUN_34 covering DRESS_21 BLUE_30 MODIFY_31 for [BLUE_36] [DRESS_29]
[!@] Construction IN_COLOR_35 covering WOMAN_0 WEAR_20 DRESS_21 BLUE_30 AGENT_22 PATIENT_23 MODIFY_31 for [ADJ_NOUN_18] 'in' [BLUE_36]
[!@] Construction BLUE_36 covering BLUE_30 for 'blue'

> Competition Traces
  IN_COLOR_35(1327) eliminated REL_SVO_WHO_27(1317)
```

```
   IN_COLOR_35(1327) eliminated WEAR_28(1317)
   IN_COLOR_35(1327) eliminated DRESS_29(1317)
   REL_SPA_WHICH_33(1310) eliminated SPA_32(339)
   ADJ_NOUN_34(1317) eliminated REL_SPA_WHICH_33(1310)
   IN_COLOR_35(1327) eliminated ADJ_NOUN_34(1317)

> Construction Structures
[ ] 138: EXIST_S_1 'there is' [WOMAN_2 'woman']
[ ] 140: EXIST_S_9 'there is' [MAN_13 'man']
[X] 138: EXIST_S_26 'there is' [DRESS_29 'dress']
[X] 339: SPA_32 [DRESS_29 'dress'] 'is' [BLUE_36 'blue']
[ ] 332: EXIST_S_1 'there is' [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']]
[ ] 732: EXIST_S_1 'there is' [IN_COLOR_35 [WOMAN_2 'woman'] 'in' [BLUE_36 'blue']]
[ ] 539: SVO_7 [WOMAN_2 'woman'] [HIT_12 'hit'] [MAN_13 'man']
[X] 327: EXIST_S_26 'there is' [REL_SPA_WHICH_33 [DRESS_29 'dress'] 'which is' [BLUE_36 'blue']]
[X] 334: EXIST_S_26 'there is' [ADJ_NOUN_34 [BLUE_36 'blue'] [DRESS_29 'dress']]
[ ] 733: SVO_7 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] [HIT_12 'hit'] [MAN_13 'man']
[ ] 1133: SVO_7 [IN_COLOR_35 [WOMAN_2 'woman'] 'in' [BLUE_36 'blue']] [HIT_12 'hit'] [MAN_13 'man']
[X] 526: EXIST_S_1 'there is' [REL_SVO_WHO_27 [WOMAN_2 'woman'] 'who' [WEAR_28 'wear'] [DRESS_29 'dress']]
[ ] 926: EXIST_S_1 'there is' [IN_COLOR_35 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] 'in' [BLUE_36 'blue']]
[X] 720: EXIST_S_1 'there is' [REL_SVO_WHO_27 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] 'who' [WEAR_28 'wear'] [DRESS_29 'dress']]
[X]  715:  EXIST_S_1  'there  is'  [REL_SVO_WHO_27  [WOMAN_2  'woman']  'who'  [WEAR_28  'wear']  [REL_SPA_WHICH_33  [DRESS_29  'dress']  'which  is'  [BLUE_36
'blue']]]
[X] 722: EXIST_S_1 'there is' [REL_SVO_WHO_27 [WOMAN_2 'woman'] 'who' [WEAR_28 'wear'] [ADJ_NOUN_34 [BLUE_36 'blue'] [DRESS_29 'dress']]]
[X] 927: SVO_7 [REL_SVO_WHO_27 [WOMAN_2 'woman'] 'who' [WEAR_28 'wear'] [DRESS_29 'dress']] [HIT_12 'hit'] [MAN_13 'man']
[*] 1327: SVO_7 [IN_COLOR_35 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] 'in' [BLUE_36 'blue']] [HIT_12 'hit'] [MAN_13 'man']
[X] 1121: SVO_7 [REL_SVO_WHO_27 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] 'who' [WEAR_28 'wear'] [DRESS_29 'dress']] [HIT_12 'hit'] [MAN_13
'man']
[X] 1116: SVO_7 [REL_SVO_WHO_27 [WOMAN_2 'woman'] 'who' [WEAR_28 'wear'] [REL_SPA_WHICH_33 [DRESS_29 'dress'] 'which is' [BLUE_36 'blue']]] [HIT_12
'hit'] [MAN_13 'man']
[X] 1123: SVO_7 [REL_SVO_WHO_27 [WOMAN_2 'woman'] 'who' [WEAR_28 'wear'] [ADJ_NOUN_34 [BLUE_36 'blue'] [DRESS_29 'dress']]] [HIT_12 'hit'] [MAN_13
'man']
[X] 909: EXIST_S_1 'there is' [REL_SVO_WHO_27 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] 'who' [WEAR_28 'wear'] [REL_SPA_WHICH_33 [DRESS_29
'dress'] 'which is' [BLUE_36 'blue']]]
[X] 916: EXIST_S_1 'there is' [REL_SVO_WHO_27 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] 'who' [WEAR_28 'wear'] [ADJ_NOUN_34 [BLUE_36 'blue']
[DRESS_29 'dress']]]
[X] 1310: SVO_7 [REL_SVO_WHO_27 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] 'who' [WEAR_28 'wear'] [REL_SPA_WHICH_33 [DRESS_29 'dress'] 'which
is' [BLUE_36 'blue']]] [HIT_12 'hit'] [MAN_13 'man']
[X] 1317: SVO_7 [REL_SVO_WHO_27 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] 'who' [WEAR_28 'wear'] [ADJ_NOUN_34 [BLUE_36 'blue'] [DRESS_29
'dress']]] [HIT_12 'hit'] [MAN_13 'man']

> Produced Utterance
  "pretty woman in blue hit man"

> Next Attention
  None

================================================================================
  Simulation Time: 7
================================================================================
> Current Attention
  None

> Schema Instances
[ x] SemRep-N WOMAN_0
[ O] Construction EXIST_S_1 covering WOMAN_0 for 'there is' [ ]
[ x] Construction WOMAN_2 covering WOMAN_0 for 'woman'
[ x] SemRep-N HIT_3
[ x] SemRep-N MAN_4
[ x] SemRep-R AGENT_5 from HIT_3 to WOMAN_0
[ x] SemRep-R PATIENT_6 from HIT_3 to MAN_4
[ x] Construction SVO_7 covering WOMAN_0 MAN_4 HIT_3 AGENT_5 PATIENT_6 for [ ] [ ] [ ]
[ O] Construction EXIST_S_9 covering MAN_4 for 'there is' [ ]
[ x] Construction HIT_12 covering HIT_3 for 'hit'
[ x] Construction MAN_13 covering MAN_4 for 'man'
[ x] SemRep-N PRETTY_14
[ x] SemRep-R MODIFY_15 from PRETTY_14 to WOMAN_0
[ x] Construction ADJ_NOUN_18 covering WOMAN_0 PRETTY_14 MODIFY_15 for [ ] [ ]
[ x] Construction PRETTY_19 covering PRETTY_14 for 'pretty'
[ x] SemRep-N WEAR_20
[ x] SemRep-N DRESS_21
[ x] SemRep-R AGENT_22 from WEAR_20 to WOMAN_0
[ x] SemRep-R PATIENT_23 from WEAR_20 to DRESS_21
[ O] Construction EXIST_S_26 covering DRESS_21 for 'there is' [ ]
[ x] SemRep-N BLUE_30
[ x] SemRep-R MODIFY_31 from BLUE_30 to DRESS_21
[ x] Construction IN_COLOR_35 covering WOMAN_0 WEAR_20 DRESS_21 BLUE_30 AGENT_22 PATIENT_23 MODIFY_31 for [ ] 'in' [ ]
[ x] Construction BLUE_36 covering BLUE_30 for 'blue'

> Next Attention
  None

================================================================================
  Simulation Time: 8
================================================================================
> Current Attention
  None

> Next Attention
  None


Simulation complete: inactivity termination.
```

The following is the simulation output for the low threshold case. Note that only the time parameter is tuned to “1” while the others are left to be infinite.

```
Template Construction Grammar (TCG) Simulator v2.5

Jinyong Lee (jinyongl@usc.edu), June 23. 2012
USC Brain Project, Computer Science Department
University of Southern California (USC)
```

```
Loading Initialization File 'TCG.ini'...
Loading Semantic Network 'TCG_semantics.txt'...
Loading Construction Vocabulary 'TCG_vocabulary.txt'...
Loading Scene 'scene_womanhitman.txt'...

Initializing Simulator...
- Max Simulation Time: 20
- Premature Production: on
- Utterance Continuity: on
- Verbal Guidance: on
- Threshold of Utterance: Time = 1, CNXs = infinite, Syllables = infinite

Beginning Simulation...

================================================================================
  Simulation Time: 1
================================================================================
> Current Attention
  None

> Next Attention
  WOMAN_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 2
================================================================================
> Current Attention
  WOMAN_AREA (perception done)

> Perceived Regions
  WOMAN_AREA

> Schema Instances
[!@] SemRep-N WOMAN_0
[!@] Construction EXIST_S_1 covering WOMAN_0 for 'there is' [WOMAN_2]
[!@] Construction WOMAN_2 covering WOMAN_0 for 'woman'

> Construction Structures
[*] 138: EXIST_S_1 'there is' [WOMAN_2 'woman']

> Produced Utterance
  "there is woman"

> Next Attention
  HIT_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 3
================================================================================
> Current Attention
  HIT_AREA (perception done)

> Perceived Regions
  HIT_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ @] Construction EXIST_S_1 covering WOMAN_0 for 'there is' [REL_SVO_WHO_10]
[ @] Construction WOMAN_2 covering WOMAN_0 for 'woman'
[!@] SemRep-N HIT_3
[!@] SemRep-N MAN_4
[!@] SemRep-R AGENT_5 from HIT_3 to WOMAN_0
[!@] SemRep-R PATIENT_6 from HIT_3 to MAN_4
[!X] Construction SVO_7 covering WOMAN_0 MAN_4 HIT_3 AGENT_5 PATIENT_6 for [WOMAN_2] [HIT_12] [MAN_13]
[!X] Construction PAS_SVO_8 covering WOMAN_0 MAN_4 HIT_3 AGENT_5 PATIENT_6 for [MAN_13] 'is' [HIT_12] '-ed by' [WOMAN_2]
[!O] Construction EXIST_S_9 covering MAN_4 for 'there is' [REL_PAS_SVO_WHO_11]
[!@] Construction REL_SVO_WHO_10 covering WOMAN_0 MAN_4 HIT_3 AGENT_5 PATIENT_6 for [WOMAN_2] 'who' [HIT_12] [MAN_13]
[!X] Construction REL_PAS_SVO_WHO_11 covering WOMAN_0 MAN_4 HIT_3 AGENT_5 PATIENT_6 for [MAN_13] 'who is' [HIT_12] '-ed by' [WOMAN_2]
[!@] Construction HIT_12 covering HIT_3 for 'hit'
[!@] Construction MAN_13 covering MAN_4 for 'man'

> Competition Traces
  SVO_7(439) eliminated PAS_SVO_8(383)
  REL_SVO_WHO_10(729) eliminated SVO_7(439)
  SVO_7(439) eliminated REL_PAS_SVO_WHO_11(423)

> Construction Structures
[ ] 140: EXIST_S_9 'there is' [MAN_13 'man']
[X] 439: SVO_7 [WOMAN_2 'woman'] [HIT_12 'hit'] [MAN_13 'man']
[X] 383: PAS_SVO_8 [MAN_13 'man'] 'is' [HIT_12 'hit'] '-ed by' [WOMAN_2 'woman']
[*] 729: EXIST_S_1 'there is' [REL_SVO_WHO_10 [WOMAN_2 'woman'] 'who' [HIT_12 'hit'] [MAN_13 'man']]
[X] 423: EXIST_S_9 'there is' [REL_PAS_SVO_WHO_11 [MAN_13 'man'] 'who is' [HIT_12 'hit'] '-ed by' [WOMAN_2 'woman']]

> Produced Utterance
  "who hit man"

> Next Attention
  WOMAN_FACE_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 4
================================================================================
> Current Attention
  WOMAN_FACE_AREA (perception done)

> Perceived Regions
  WOMAN_FACE_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ x] Construction EXIST_S_1 covering WOMAN_0 for 'there is' [ADJ_NOUN_18]
[ @] Construction WOMAN_2 covering WOMAN_0 for 'woman'
[ x] SemRep-N HIT_3
[ x] SemRep-N MAN_4
[ x] SemRep-R AGENT_5 from HIT_3 to WOMAN_0
[ x] SemRep-R PATIENT_6 from HIT_3 to MAN_4
[ O] Construction EXIST_S_9 covering MAN_4 for 'there is' [ ]
```

```
[ x] Construction REL_SVO_WHO_10 covering WOMAN_0 MAN_4 HIT_3 AGENT_5 PATIENT_6 for [WOMAN_2] 'who' [HIT_12] [MAN_13]
[ x] Construction HIT_12 covering HIT_3 for 'hit'
[ x] Construction MAN_13 covering MAN_4 for 'man'
[!@] SemRep-N PRETTY_14
[!@] SemRep-R MODIFY_15 from PRETTY_14 to WOMAN_0
[!@] Construction SPA_16 covering WOMAN_0 PRETTY_14 MODIFY_15 for [WOMAN_2] 'is' [PRETTY_19]
[!X] Construction REL_SPA_WHO_17 covering WOMAN_0 PRETTY_14 MODIFY_15 for [WOMAN_2] 'who is' [PRETTY_19]
[!X] Construction ADJ_NOUN_18 covering WOMAN_0 PRETTY_14 MODIFY_15 for [PRETTY_19] [WOMAN_2]
[!@] Construction PRETTY_19 covering PRETTY_14 for 'pretty'

> Competition Traces
  SPA_16(237) eliminated REL_SPA_WHO_17(127)
  SPA_16(237) eliminated ADJ_NOUN_18(132)

> Construction Structures
[*] 237: SPA_16 [WOMAN_2 'woman'] 'is' [PRETTY_19 'pretty']
[X] 127: EXIST_S_1 'there is' [REL_SPA_WHO_17 [WOMAN_2 'woman'] 'who is' [PRETTY_19 'pretty']]
[X] 132: EXIST_S_1 'there is' [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']]
[X] -77: EXIST_S_1 'there is' [REL_SVO_WHO_10 [ADJ_NOUN_18 [PRETTY_19 'pretty'] [WOMAN_2 'woman']] 'who' [HIT_12 'hit'] [MAN_13 'man']]
[ ] 28: SPA_16 [REL_SVO_WHO_10 [WOMAN_2 'woman'] 'who' [HIT_12 'hit'] [MAN_13 'man']] 'is' [PRETTY_19 'pretty']

> Produced Utterance
  "woman is pretty"

> Next Attention
  DRESS_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 5
================================================================================
> Current Attention
  DRESS_AREA (perception done)

> Perceived Regions
  DRESS_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ @] Construction WOMAN_2 covering WOMAN_0 for 'woman'
[ x] SemRep-N PRETTY_14
[ x] SemRep-R MODIFY_15 from PRETTY_14 to WOMAN_0
[ x] Construction SPA_16 covering WOMAN_0 PRETTY_14 MODIFY_15 for [REL_SVO_WHO_27] 'is' [PRETTY_19]
[ x] Construction PRETTY_19 covering PRETTY_14 for 'pretty'
[!@] SemRep-N WEAR_20
[!@] SemRep-N DRESS_21
[!@] SemRep-R AGENT_22 from WEAR_20 to WOMAN_0
[!@] SemRep-R PATIENT_23 from WEAR_20 to DRESS_21
[!@] Construction SVO_24 covering WOMAN_0 DRESS_21 WEAR_20 AGENT_22 PATIENT_23 for [WOMAN_2] [WEAR_28] [DRESS_29]
[!X] Construction PAS_SVO_25 covering WOMAN_0 DRESS_21 WEAR_20 AGENT_22 PATIENT_23 for [DRESS_29] 'is' [WEAR_28] '-ed by' [WOMAN_2]
[!O] Construction EXIST_S_26 covering DRESS_21 for 'there is' [DRESS_29]
[!X] Construction REL_SVO_WHO_27 covering WOMAN_0 DRESS_21 WEAR_20 AGENT_22 PATIENT_23 for [WOMAN_2] 'who' [WEAR_28] [DRESS_29]
[!@] Construction WEAR_28 covering WEAR_20 for 'wear'
[!@] Construction DRESS_29 covering DRESS_21 for 'dress'

> Competition Traces
  SVO_24(436) eliminated PAS_SVO_25(380)
  SVO_24(436) eliminated REL_SVO_WHO_27(325)

> Construction Structures
[ ] 138: EXIST_S_26 'there is' [DRESS_29 'dress']
[*] 436: SVO_24 [WOMAN_2 'woman'] [WEAR_28 'wear'] [DRESS_29 'dress']
[X] 380: PAS_SVO_25 [DRESS_29 'dress'] 'is' [WEAR_28 'wear'] '-ed by' [WOMAN_2 'woman']
[X] 325: SPA_16 [REL_SVO_WHO_27 [WOMAN_2 'woman'] 'who' [WEAR_28 'wear'] [DRESS_29 'dress']] 'is' [PRETTY_19 'pretty']

> Produced Utterance
  "woman wear dress"

> Next Attention
  DRESS_FOCUS_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 6
================================================================================
> Current Attention
  DRESS_FOCUS_AREA (perception done)

> Perceived Regions
  DRESS_FOCUS_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ @] Construction WOMAN_2 covering WOMAN_0 for 'woman'
[ @] SemRep-N WEAR_20
[ @] SemRep-N DRESS_21
[ @] SemRep-R AGENT_22 from WEAR_20 to WOMAN_0
[ @] SemRep-R PATIENT_23 from WEAR_20 to DRESS_21
[ @] Construction SVO_24 covering WOMAN_0 DRESS_21 WEAR_20 AGENT_22 PATIENT_23 for [WOMAN_2] [WEAR_28] [REL_SPA_WHICH_33]
[ O] Construction EXIST_S_26 covering DRESS_21 for 'there is' [ADJ_NOUN_34]
[ @] Construction WEAR_28 covering WEAR_20 for 'wear'
[ @] Construction DRESS_29 covering DRESS_21 for 'dress'
[!@] SemRep-N BLUE_30
[!@] SemRep-R MODIFY_31 from BLUE_30 to DRESS_21
[!X] Construction SPA_32 covering DRESS_21 BLUE_30 MODIFY_31 for [DRESS_29] 'is' [BLUE_36]
[!@] Construction REL_SPA_WHICH_33 covering DRESS_21 BLUE_30 MODIFY_31 for [DRESS_29] 'which is' [BLUE_36]
[!X] Construction ADJ_NOUN_34 covering DRESS_21 BLUE_30 MODIFY_31 for [BLUE_36] [DRESS_29]
[!X] Construction IN_COLOR_35 covering WOMAN_0 WEAR_20 DRESS_21 BLUE_30 AGENT_22 PATIENT_23 MODIFY_31 for [WOMAN_2] 'in' [BLUE_36]
[!@] Construction BLUE_36 covering BLUE_30 for 'blue'

> Competition Traces
  SVO_24(1325) eliminated IN_COLOR_35(589)
  REL_SPA_WHICH_33(1325) eliminated SPA_32(239)
  SPA_32(239) eliminated ADJ_NOUN_34(234)

> Construction Structures
[X] 589: IN_COLOR_35 [WOMAN_2 'woman'] 'in' [BLUE_36 'blue']
[X] 239: SPA_32 [DRESS_29 'dress'] 'is' [BLUE_36 'blue']
[ ] 227: EXIST_S_26 'there is' [REL_SPA_WHICH_33 [DRESS_29 'dress'] 'which is' [BLUE_36 'blue']]
```

```
[X] 234: EXIST_S_26 'there is' [ADJ_NOUN_34 [BLUE_36 'blue'] [DRESS_29 'dress']]
[*] 1325: SVO_24 [WOMAN_2 'woman'] [WEAR_28 'wear'] [REL_SPA_WHICH_33 [DRESS_29 'dress'] 'which is' [BLUE_36 'blue']]
[X] 132: SVO_24 [WOMAN_2 'woman'] [WEAR_28 'wear'] [ADJ_NOUN_34 [BLUE_36 'blue'] [DRESS_29 'dress']]

> Produced Utterance
  "which is blue"

> Next Attention
  None

================================================================================
  Simulation Time: 7
================================================================================
> Current Attention
  None

> Schema Instances
[ x] SemRep-N WOMAN_0
[ x] Construction WOMAN_2 covering WOMAN_0 for 'woman'
[ x] SemRep-N WEAR_20
[ x] SemRep-N DRESS_21
[ x] SemRep-R AGENT_22 from WEAR_20 to WOMAN_0
[ x] SemRep-R PATIENT_23 from WEAR_20 to DRESS_21
[ x] Construction SVO_24 covering WOMAN_0 DRESS_21 WEAR_20 AGENT_22 PATIENT_23 for [ ] [ ] [ ]
[ O] Construction EXIST_S_26 covering DRESS_21 for 'there is' [ ]
[ x] Construction WEAR_28 covering WEAR_20 for 'wear'
[ x] Construction DRESS_29 covering DRESS_21 for 'dress'
[ x] SemRep-N BLUE_30
[ x] SemRep-R MODIFY_31 from BLUE_30 to DRESS_21
[ x] Construction REL_SPA_WHICH_33 covering DRESS_21 BLUE_30 MODIFY_31 for [ ] 'which is' [ ]
[ x] Construction BLUE_36 covering BLUE_30 for 'blue'

> Next Attention
  None

================================================================================
  Simulation Time: 8
================================================================================
> Current Attention
  None

> Next Attention
  None


Simulation complete: inactivity termination.
```

Appendix D. Simulation Demo

This appendix provides the entire simulation result as well as the scene description file presented in Section 4.6. The following is the scene description file used for the simulation.

```
#
# TCG Scene: Woman-hit-man
#
# layout of the hitting event is given first
#


image: "woman hits man.jpg"
resolution: 400 * 326


# scene gist
region GIST
{
            location: 216, 117 size: 150, 100
            saliency: 100                       # saliency doesn't matter
            uncertainty: 0                      # instantly perceived

            # layout
            perceive WOMAN=ENTITY, HIT=ACTION, MAN=ENTITY
            perceive HIT_AGENT, HIT_PATIENT
}

region MAN_AREA
{
            location: 164, 204 size: 80, 180
            saliency: 100                       # most salient region: possibly fixated first
            uncertainty: 1

            # associated perceptual schema
            object MAN { concept: MAN }

            perceive MAN
}

region HIT_AREA
{
            location: 213, 110 size: 60, 20
            saliency: 90
            uncertainty: 1

            object HIT { concept: HIT }
            relation HIT_AGENT { concept: AGENT from: HIT to: WOMAN }
            relation HIT_PATIENT { concept: PATIENT from: HIT to: MAN }

            perceive HIT, HIT_AGENT, HIT_PATIENT
}
```

```
region WOMAN_AREA
{
            location: 283, 205 size: 50, 150
            saliency: 70
            uncertainty: 1

            object WOMAN { concept: WOMAN }

            perceive WOMAN
}

region MAN_FACE_AREA
{
            location: 164, 204 size: 80, 180
            saliency: 50
            uncertainty: 1

            object HANDSOME { concept: HANDSOME }
            relation HANDSOME_MODIFY { concept: MODIFY from: HANDSOME to: MAN }

            perceive HANDSOME, HANDSOME_MODIFY
}
```

The following is the simulation output, which is a type of low threshold case (the time parameter is set to 1, and the number of construction instances and syllables are set to infinite).

```
Template Construction Grammar (TCG) Simulator v2.5

Jinyong Lee (jinyongl@usc.edu), June 23. 2012
USC Brain Project, Computer Science Department
University of Southern California (USC)

Loading Initialization File 'TCG.ini'...
Loading Semantic Network 'TCG_semantics.txt'...
Loading Construction Vocabulary 'TCG_vocabulary.txt'...
Loading Scene 'scene_demo.txt'...

Initializing Simulator...
- Max Simulation Time: 20
- Premature Production: on
- Utterance Continuity: on
- Verbal Guidance: on
- Threshold of Utterance: Time = 1, CNXs = infinite, Syllables = infinite

Beginning Simulation...

================================================================================
  Simulation Time: 1
================================================================================
> Current Attention
  None

> Perceived Regions
  GIST

> Schema Instances
[!O] SemRep-N ENTITY_0
[!O] SemRep-N ACTION_1
[!O] SemRep-N ENTITY_2
[!@] SemRep-R AGENT_3 from ACTION_1 to ENTITY_0
[!@] SemRep-R PATIENT_4 from ACTION_1 to ENTITY_2
[!@] Construction SVO_5 covering ENTITY_0 ENTITY_2 ACTION_1 AGENT_3 PATIENT_4 for [ ] [ ] [ ]
[!X] Construction PAS_SVO_6 covering ENTITY_0 ENTITY_2 ACTION_1 AGENT_3 PATIENT_4 for [ ] 'is' [ ] '-ed by' [ ]

> Competition Traces
  SVO_5(250) eliminated PAS_SVO_6(194)

> Construction Structures
[*] 250: SVO_5 [ ] [ ] [ ]
[X] 194: PAS_SVO_6 [ ] 'is' [ ] '-ed by' [ ]

> Produced Utterance
  "uh..."

> Next Attention
  WOMAN_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 2
================================================================================
> Current Attention
  WOMAN_AREA (perception done)

> Perceived Regions
  WOMAN_AREA

> Schema Instances
[!@] SemRep-N WOMAN_0
[ O] SemRep-N ACTION_1
[ O] SemRep-N ENTITY_2
[ @] SemRep-R AGENT_3 from ACTION_1 to WOMAN_0
[ @] SemRep-R PATIENT_4 from ACTION_1 to ENTITY_2
[ @] Construction SVO_5 covering WOMAN_0 ENTITY_2 ACTION_1 AGENT_3 PATIENT_4 for [WOMAN_11] [ ] [ ]
[!X] Construction PAS_SVO_8 covering WOMAN_0 ENTITY_2 ACTION_1 AGENT_3 PATIENT_4 for [ ] 'is' [ ] '-ed by' [WOMAN_11]
[!O] Construction EXIST_S_9 covering WOMAN_0 for 'there is' [REL_SVO_WHO_10]
[!X] Construction REL_SVO_WHO_10 covering WOMAN_0 ENTITY_2 ACTION_1 AGENT_3 PATIENT_4 for [WOMAN_11] 'who' [ ] [ ]
[!@] Construction WOMAN_11 covering WOMAN_0 for 'woman'

> Competition Traces
  SVO_5(445) eliminated PAS_SVO_8(289)
  SVO_5(445) eliminated REL_SVO_WHO_10(335)
```

```
> Construction Structures
[*] 445: SVO_5 [WOMAN_11 'woman'] [ ] [ ]
[X] 289: PAS_SVO_8 [ ] 'is' [ ] '-ed by' [WOMAN_11 'woman']
[ ] 138: EXIST_S_9 'there is' [WOMAN_11 'woman']
[X] 335: EXIST_S_9 'there is' [REL_SVO_WHO_10 [WOMAN_11 'woman'] 'who' [ ] [ ]]

> Produced Utterance
  "woman..."

> Next Attention
  HIT_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 3
================================================================================
> Current Attention
  HIT_AREA (perception done)

> Perceived Regions
  HIT_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[!@] SemRep-N HIT_1
[ O] SemRep-N ENTITY_2
[!@] SemRep-R AGENT_3 from HIT_1 to WOMAN_0
[!@] SemRep-R PATIENT_4 from HIT_1 to ENTITY_2
[ @] Construction SVO_5 covering WOMAN_0 ENTITY_2 HIT_1 AGENT_3 PATIENT_4 for [WOMAN_11] [HIT_15] [ ]
[ O] Construction EXIST_S_9 covering WOMAN_0 for 'there is' [REL_SVO_WHO_14]
[ @] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[!X] Construction PAS_SVO_13 covering WOMAN_0 ENTITY_2 HIT_1 AGENT_3 PATIENT_4 for [ ] 'is' [HIT_15] '-ed by' [WOMAN_11]
[!X] Construction REL_SVO_WHO_14 covering WOMAN_0 ENTITY_2 HIT_1 AGENT_3 PATIENT_4 for [WOMAN_11] 'who' [HIT_15] [ ]
[!@] Construction HIT_15 covering HIT_1 for 'hit'

> Competition Traces
  SVO_5(742) eliminated PAS_SVO_13(286)
  SVO_5(742) eliminated REL_SVO_WHO_14(332)

> Construction Structures
[*] 742: SVO_5 [WOMAN_11 'woman'] [HIT_15 'hit'] [ ]
[X] 286: PAS_SVO_13 [ ] 'is' [HIT_15 'hit'] '-ed by' [WOMAN_11 'woman']
[X] 332: EXIST_S_9 'there is' [REL_SVO_WHO_14 [WOMAN_11 'woman'] 'who' [HIT_15 'hit'] [ ]]

> Produced Utterance
  "hit..."

> Next Attention
  MAN_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 4
================================================================================
> Current Attention
  MAN_AREA (perception done)

> Perceived Regions
  MAN_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ @] SemRep-N HIT_1
[!@] SemRep-N MAN_2
[ @] SemRep-R AGENT_3 from HIT_1 to WOMAN_0
[ @] SemRep-R PATIENT_4 from HIT_1 to MAN_2
[ @] Construction SVO_5 covering WOMAN_0 MAN_2 HIT_1 AGENT_3 PATIENT_4 for [WOMAN_11] [HIT_15] [MAN_21]
[ O] Construction EXIST_S_9 covering WOMAN_0 for 'there is' [REL_SVO_WHO_19]
[ @] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ @] Construction HIT_15 covering HIT_1 for 'hit'
[!X] Construction PAS_SVO_17 covering WOMAN_0 MAN_2 HIT_1 AGENT_3 PATIENT_4 for [MAN_21] 'is' [HIT_15] '-ed by' [WOMAN_11]
[!O] Construction EXIST_S_18 covering MAN_2 for 'there is' [REL_PAS_SVO_WHO_20]
[!X] Construction REL_SVO_WHO_19 covering WOMAN_0 MAN_2 HIT_1 AGENT_3 PATIENT_4 for [WOMAN_11] 'who' [HIT_15] [MAN_21]
[!X] Construction REL_PAS_SVO_WHO_20 covering WOMAN_0 MAN_2 HIT_1 AGENT_3 PATIENT_4 for [MAN_21] 'who is' [HIT_15] '-ed by' [WOMAN_11]
[!@] Construction MAN_21 covering MAN_2 for 'man'

> Competition Traces
  SVO_5(1039) eliminated PAS_SVO_17(283)
  SVO_5(1039) eliminated REL_SVO_WHO_19(329)
  SVO_5(1039) eliminated REL_PAS_SVO_WHO_20(323)

> Construction Structures
[ ] 140: EXIST_S_18 'there is' [MAN_21 'man']
[*] 1039: SVO_5 [WOMAN_11 'woman'] [HIT_15 'hit'] [MAN_21 'man']
[X] 283: PAS_SVO_17 [MAN_21 'man'] 'is' [HIT_15 'hit'] '-ed by' [WOMAN_11 'woman']
[X] 329: EXIST_S_9 'there is' [REL_SVO_WHO_19 [WOMAN_11 'woman'] 'who' [HIT_15 'hit'] [MAN_21 'man']]
[X] 323: EXIST_S_18 'there is' [REL_PAS_SVO_WHO_20 [MAN_21 'man'] 'who is' [HIT_15 'hit'] '-ed by' [WOMAN_11 'woman']]

> Produced Utterance
  "man"

> Next Attention
  MAN_FACE_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 5
================================================================================
> Current Attention
  MAN_FACE_AREA (perception done)

> Perceived Regions
  MAN_FACE_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ @] SemRep-N HIT_1
[ @] SemRep-N MAN_2
[ @] SemRep-R AGENT_3 from HIT_1 to WOMAN_0
[ @] SemRep-R PATIENT_4 from HIT_1 to MAN_2
```

```
[ @] Construction SVO_5 covering WOMAN_0 MAN_2 HIT_1 AGENT_3 PATIENT_4 for [WOMAN_11] [HIT_15] [REL_SPA_WHO_25]
[ O] Construction EXIST_S_9 covering WOMAN_0 for 'there is' [ ]
[ @] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ @] Construction HIT_15 covering HIT_1 for 'hit'
[ O] Construction EXIST_S_18 covering MAN_2 for 'there is' [ADJ_NOUN_26]
[ @] Construction MAN_21 covering MAN_2 for 'man'
[!@] SemRep-N HANDSOME_22
[!@] SemRep-R MODIFY_23 from HANDSOME_22 to MAN_2
[!X] Construction SPA_24 covering MAN_2 HANDSOME_22 MODIFY_23 for [MAN_21] 'is' [HANDSOME_27]
[!@] Construction REL_SPA_WHO_25 covering MAN_2 HANDSOME_22 MODIFY_23 for [MAN_21] 'who is' [HANDSOME_27]
[!X] Construction ADJ_NOUN_26 covering MAN_2 HANDSOME_22 MODIFY_23 for [HANDSOME_27] [MAN_21]
[!@] Construction HANDSOME_27 covering HANDSOME_22 for 'handsome'

> Competition Traces
  REL_SPA_WHO_25(1326) eliminated SPA_24(237)
  SPA_24(237) eliminated ADJ_NOUN_26(232)

> Construction Structures
[X] 237: SPA_24 [MAN_21 'man'] 'is' [HANDSOME_27 'handsome']
[ ] 227: EXIST_S_18 'there is' [REL_SPA_WHO_25 [MAN_21 'man'] 'who is' [HANDSOME_27 'handsome']]
[X] 232: EXIST_S_18 'there is' [ADJ_NOUN_26 [HANDSOME_27 'handsome'] [MAN_21 'man']]
[*] 1326: SVO_5 [WOMAN_11 'woman'] [HIT_15 'hit'] [REL_SPA_WHO_25 [MAN_21 'man'] 'who is' [HANDSOME_27 'handsome']]
[X] 131: SVO_5 [WOMAN_11 'woman'] [HIT_15 'hit'] [ADJ_NOUN_26 [HANDSOME_27 'handsome'] [MAN_21 'man']]

> Produced Utterance
  "who is handsome"

> Next Attention
  None

================================================================================
  Simulation Time: 6
================================================================================
> Current Attention
  None

> Schema Instances
[ x] SemRep-N WOMAN_0
[ x] SemRep-N HIT_1
[ x] SemRep-N MAN_2
[ x] SemRep-R AGENT_3 from HIT_1 to WOMAN_0
[ x] SemRep-R PATIENT_4 from HIT_1 to MAN_2
[ x] Construction SVO_5 covering WOMAN_0 MAN_2 HIT_1 AGENT_3 PATIENT_4 for [ ] [ ] [ ] [ ]
[ O] Construction EXIST_S_9 covering WOMAN_0 for 'there is' [ ]
[ x] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ x] Construction HIT_15 covering HIT_1 for 'hit'
[ O] Construction EXIST_S_18 covering MAN_2 for 'there is' [ ]
[ x] Construction MAN_21 covering MAN_2 for 'man'
[ x] SemRep-N HANDSOME_22
[ x] SemRep-R MODIFY_23 from HANDSOME_22 to MAN_2
[ x] Construction REL_SPA_WHO_25 covering MAN_2 HANDSOME_22 MODIFY_23 for [ ] 'who is' [ ]
[ x] Construction HANDSOME_27 covering HANDSOME_22 for 'handsome'

> Next Attention
  None

================================================================================
  Simulation Time: 7
================================================================================
> Current Attention
  None

> Next Attention
  None


Simulation complete: inactivity termination.
```

Appendix E. Simulation Result of High and Low Threshold (Cholitas Scene)

In Section 4.6, visualized illustrations of the simulation of high and low threshold cases are presented. This appendix provides the actual simulation results corresponding to the illustrations.

The following is the scene description file used for the simulation, which is based on the scene (Cholitas scene) used in the eye-tracking experiment.

```
#
# TCG Scene: Cholita scene
#

image: "cholitas.jpg"
resolution: 1024 * 768


region LEFT_WOMAN_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 100
            uncertainty: 1

            object WOMAN_L { concept: WOMAN }
            object DRESS_L { concept: DRESS }
            object WEAR_L { concept: WEAR }
            relation WEAR_AGENT_L { concept: AGENT from: WEAR_L to: WOMAN_L }
            relation WEAR_PATIENT_L { concept: PATIENT from: WEAR_L to: DRESS_L }
```

```
            perceive WOMAN_L, WEAR_L, DRESS_L, WEAR_AGENT_L, WEAR_PATIENT_L
}

region LEFT_DRESS_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 90
            uncertainty: 1

            object GREEN { concept: GREEN }
            relation GREEN_MODIFY { concept: MODIFY from: GREEN to: DRESS_L }

            perceive GREEN, GREEN_MODIFY
}

region KICK_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 80
            uncertainty: 1

            object KICK { concept: KICK }
            relation KICK_AGENT { concept: AGENT from: KICK to: WOMAN_L }
            relation KICK_PATIENT { concept: PATIENT from: KICK to: WOMAN_R }

            perceive KICK, KICK_PATIENT, KICK_AGENT
            perceive WOMAN_R = HUMAN
}

region RIGHT_WOMAN_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 70
            uncertainty: 1

            object WOMAN_R { concept: WOMAN }
            object DRESS_R { concept: DRESS }
            object WEAR_R { concept: WEAR }
            relation WEAR_AGENT_R { concept: AGENT from: WEAR_R to: WOMAN_R }
            relation WEAR_PATIENT_R { concept: PATIENT from: WEAR_R to: DRESS_R }

            perceive WOMAN_R, WEAR_R, DRESS_R, WEAR_AGENT_R, WEAR_PATIENT_R
}

region RIGHT_DRESS_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 60
            uncertainty: 1

            object BLUE { concept: BLUE }
            relation BLUE_MODIFY { concept: MODIFY from: BLUE to: DRESS_R }

            perceive BLUE, BLUE_MODIFY
}

region BOXINGRING_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 50
            uncertainty: 2                      # assume that it takes a bit longer to figure out

            object BOXINGRING { concept: BOXINGRING }
            relation IN_BOXINGRING { concept: IN from: KICK to: BOXINGRING }

            perceive BOXINGRING, IN_BOXINGRING
}

region PEOPLE_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 40
            uncertainty: 1

            object PEOPLE { concept: PEOPLE }

            perceive PEOPLE
}

region PEOPLE_FOCUS_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 30
            uncertainty: 1

            object WATCH { concept: WATCH }
            relation WATCH_AGENT { concept: AGENT from: WATCH to: PEOPLE }
            relation WHILE { concept: CONCURRENT from: KICK to: WATCH }

            perceive WATCH, WATCH_AGENT, WHILE
}
```

The following is the simulation output for the high threshold case. In this case, all of the threshold parameters are set to infinite.

```
Template Construction Grammar (TCG) Simulator v2.5

Jinyong Lee (jinyongl@usc.edu), June 23. 2012
USC Brain Project, Computer Science Department
University of Southern California (USC)

Loading Initialization File 'TCG.ini'...
Loading Semantic Network 'TCG_semantics.txt'...
```

```
Loading Construction Vocabulary 'TCG_vocabulary.txt'...
Loading Scene 'scene_cholita.txt'...

Initializing Simulator...
- Max Simulation Time: 20
- Premature Production: on
- Utterance Continuity: on
- Verbal Guidance: on
- Threshold of Utterance: Time = infinite, CNXs = infinite, Syllables = infinite

Beginning Simulation...

================================================================================
  Simulation Time: 1
================================================================================
> Current Attention
  None

> Next Attention
  LEFT_WOMAN_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 2
================================================================================
> Current Attention
  LEFT_WOMAN_AREA (perception done)

> Perceived Regions
  LEFT_WOMAN_AREA

> Schema Instances
[!O] SemRep-N WOMAN_0
[!O] SemRep-N WEAR_1
[!O] SemRep-N DRESS_2
[!O] SemRep-R AGENT_3 from WEAR_1 to WOMAN_0
[!O] SemRep-R PATIENT_4 from WEAR_1 to DRESS_2
[!O] Construction SVO_5 covering WOMAN_0 DRESS_2 WEAR_1 AGENT_3 PATIENT_4 for [WOMAN_11] [WEAR_10] [DRESS_12]
[!X] Construction PAS_SVO_6 covering WOMAN_0 DRESS_2 WEAR_1 AGENT_3 PATIENT_4 for [DRESS_12] 'is' [WEAR_10] '-ed by' [WOMAN_11]
[!O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [REL_SVO_WHO_9]
[!O] Construction EXIST_S_8 covering DRESS_2 for 'there is' [DRESS_12]
[!X] Construction REL_SVO_WHO_9 covering WOMAN_0 DRESS_2 WEAR_1 AGENT_3 PATIENT_4 for [WOMAN_11] 'who' [WEAR_10] [DRESS_12]
[!O] Construction WEAR_10 covering WEAR_1 for 'wear'
[!O] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[!O] Construction DRESS_12 covering DRESS_2 for 'dress'

> Competition Traces
  SVO_5(536) eliminated PAS_SVO_6(480)
  SVO_5(536) eliminated REL_SVO_WHO_9(526)

> Construction Structures
[ ] 138: EXIST_S_7 'there is' [WOMAN_11 'woman']
[ ] 138: EXIST_S_8 'there is' [DRESS_12 'dress']
[ ] 536: SVO_5 [WOMAN_11 'woman'] [WEAR_10 'wear'] [DRESS_12 'dress']
[X] 480: PAS_SVO_6 [DRESS_12 'dress'] 'is' [WEAR_10 'wear'] '-ed by' [WOMAN_11 'woman']
[X] 526: EXIST_S_7 'there is' [REL_SVO_WHO_9 [WOMAN_11 'woman'] 'who' [WEAR_10 'wear'] [DRESS_12 'dress']]

> Next Attention
  LEFT_DRESS_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 3
================================================================================
> Current Attention
  LEFT_DRESS_AREA (perception done)

> Perceived Regions
  LEFT_DRESS_AREA

> Schema Instances
[ O] SemRep-N WOMAN_0
[ O] SemRep-N WEAR_1
[ O] SemRep-N DRESS_2
[ O] SemRep-R AGENT_3 from WEAR_1 to WOMAN_0
[ O] SemRep-R PATIENT_4 from WEAR_1 to DRESS_2
[ O] Construction SVO_5 covering WOMAN_0 DRESS_2 WEAR_1 AGENT_3 PATIENT_4 for [WOMAN_11] [WEAR_10] [ADJ_NOUN_17]
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [IN_COLOR_18]
[ O] Construction EXIST_S_8 covering DRESS_2 for 'there is' [ADJ_NOUN_17]
[ O] Construction WEAR_10 covering WEAR_1 for 'wear'
[ O] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ O] Construction DRESS_12 covering DRESS_2 for 'dress'
[!O] SemRep-N GREEN_13
[!O] SemRep-R MODIFY_14 from GREEN_13 to DRESS_2
[!X] Construction SPA_15 covering DRESS_2 GREEN_13 MODIFY_14 for [DRESS_12] 'is' [GREEN_19]
[!X] Construction REL_SPA_WHICH_16 covering DRESS_2 GREEN_13 MODIFY_14 for [DRESS_12] 'which is' [GREEN_19]
[!O] Construction ADJ_NOUN_17 covering DRESS_2 GREEN_13 MODIFY_14 for [GREEN_19] [DRESS_12]
[!O] Construction IN_COLOR_18 covering WOMAN_0 WEAR_1 DRESS_2 GREEN_13 AGENT_3 PATIENT_4 MODIFY_14 for [WOMAN_11] 'in' [GREEN_19]
[!O] Construction GREEN_19 covering GREEN_13 for 'green'

> Competition Traces
  REL_SPA_WHICH_16(724) eliminated SPA_15(338)
  ADJ_NOUN_17(731) eliminated REL_SPA_WHICH_16(724)

> Construction Structures
[ ] 138: EXIST_S_7 'there is' [WOMAN_11 'woman']
[ ] 138: EXIST_S_8 'there is' [DRESS_12 'dress']
[X] 338: SPA_15 [DRESS_12 'dress'] 'is' [GREEN_19 'green']
[ ] 731: EXIST_S_7 'there is' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]
[X] 326: EXIST_S_8 'there is' [REL_SPA_WHICH_16 [DRESS_12 'dress'] 'which is' [GREEN_19 'green']]
[ ] 333: EXIST_S_8 'there is' [ADJ_NOUN_17 [GREEN_19 'green'] [DRESS_12 'dress']]
[ ] 536: SVO_5 [WOMAN_11 'woman'] [WEAR_10 'wear'] [DRESS_12 'dress']
[X] 724: SVO_5 [WOMAN_11 'woman'] [WEAR_10 'wear'] [REL_SPA_WHICH_16 [DRESS_12 'dress'] 'which is' [GREEN_19 'green']]
[ ] 731: SVO_5 [WOMAN_11 'woman'] [WEAR_10 'wear'] [ADJ_NOUN_17 [GREEN_19 'green'] [DRESS_12 'dress']]

> Next Attention
  KICK_AREA (uncertainty left: 1)

================================================================================
```

```
  Simulation Time: 4
================================================================================
> Current Attention
  KICK_AREA (perception done)

> Perceived Regions
  KICK_AREA

> Schema Instances
[ O] SemRep-N WOMAN_0
[ O] SemRep-N WEAR_1
[ O] SemRep-N DRESS_2
[ O] SemRep-R AGENT_3 from WEAR_1 to WOMAN_0
[ O] SemRep-R PATIENT_4 from WEAR_1 to DRESS_2
[ X] Construction SVO_5 covering WOMAN_0 DRESS_2 WEAR_1 AGENT_3 PATIENT_4 for [REL_SVO_WHO_27] [WEAR_10] [ADJ_NOUN_17]
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [REL_SVO_WHO_27]
[ O] Construction EXIST_S_8 covering DRESS_2 for 'there is' [ADJ_NOUN_17]
[ X] Construction WEAR_10 covering WEAR_1 for 'wear'
[ O] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ X] Construction DRESS_12 covering DRESS_2 for 'dress'
[ O] SemRep-N GREEN_13
[ O] SemRep-R MODIFY_14 from GREEN_13 to DRESS_2
[ X] Construction ADJ_NOUN_17 covering DRESS_2 GREEN_13 MODIFY_14 for [GREEN_19] [DRESS_12]
[ O] Construction IN_COLOR_18 covering WOMAN_0 WEAR_1 DRESS_2 GREEN_13 AGENT_3 PATIENT_4 MODIFY_14 for [WOMAN_11] 'in' [GREEN_19]
[ O] Construction GREEN_19 covering GREEN_13 for 'green'
[!O] SemRep-N KICK_20
[!O] SemRep-R PATIENT_21 from KICK_20 to HUMAN_23
[!O] SemRep-R AGENT_22 from KICK_20 to WOMAN_0
[!O] SemRep-N HUMAN_23
[!O] Construction SVO_24 covering WOMAN_0 HUMAN_23 KICK_20 AGENT_22 PATIENT_21 for [IN_COLOR_18] [KICK_29] [ ]
[!X] Construction PAS_SVO_25 covering WOMAN_0 HUMAN_23 KICK_20 AGENT_22 PATIENT_21 for [ ] 'is' [KICK_29] '-ed by' [IN_COLOR_18]
[!O] Construction EXIST_S_26 covering HUMAN_23 for 'there is' [REL_PAS_SVO_WHO_28]
[!X] Construction REL_SVO_WHO_27 covering WOMAN_0 HUMAN_23 KICK_20 AGENT_22 PATIENT_21 for [WOMAN_11] 'who' [KICK_29] [ ]
[!X] Construction REL_PAS_SVO_WHO_28 covering WOMAN_0 HUMAN_23 KICK_20 AGENT_22 PATIENT_21 for [ ] 'who is' [KICK_29] '-ed by' [IN_COLOR_18]
[!O] Construction KICK_29 covering KICK_20 for 'kick'

> Competition Traces
  IN_COLOR_18(1034) eliminated SVO_5(1024)
  IN_COLOR_18(1034) eliminated WEAR_10(1024)
  IN_COLOR_18(1034) eliminated DRESS_12(1024)
  IN_COLOR_18(1034) eliminated ADJ_NOUN_17(1024)
  SVO_24(1034) eliminated PAS_SVO_25(978)
  SVO_24(1034) eliminated REL_SVO_WHO_27(1024)
  SVO_24(1034) eliminated REL_PAS_SVO_WHO_28(1018)

> Construction Structures
[ ] 138: EXIST_S_7 'there is' [WOMAN_11 'woman']
[X] 138: EXIST_S_8 'there is' [DRESS_12 'dress']
[ ] 441: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [ ]
[X] 385: PAS_SVO_25 [ ] 'is' [KICK_29 'kick'] '-ed by' [WOMAN_11 'woman']
[ ] 731: EXIST_S_7 'there is' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]
[X] 431: EXIST_S_7 'there is' [REL_SVO_WHO_27 [WOMAN_11 'woman'] 'who' [KICK_29 'kick'] [ ]]
[X] 333: EXIST_S_8 'there is' [ADJ_NOUN_17 [GREEN_19 'green'] [DRESS_12 'dress']]
[X] 536: SVO_5 [WOMAN_11 'woman'] [WEAR_10 'wear'] [DRESS_12 'dress']
[X] 425: EXIST_S_26 'there is' [REL_PAS_SVO_WHO_28 [ ] 'who is' [KICK_29 'kick'] '-ed by' [WOMAN_11 'woman']]
[ ] 1034: SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [ ]
[X] 978: PAS_SVO_25 [ ] 'is' [KICK_29 'kick'] '-ed by' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]
[X] 731: SVO_5 [WOMAN_11 'woman'] [WEAR_10 'wear'] [ADJ_NOUN_17 [GREEN_19 'green'] [DRESS_12 'dress']]
[X] 829: SVO_5 [REL_SVO_WHO_27 [WOMAN_11 'woman'] 'who' [KICK_29 'kick'] [ ]] [WEAR_10 'wear'] [DRESS_12 'dress']
[X] 1024: EXIST_S_7 'there is' [REL_SVO_WHO_27 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] 'who' [KICK_29 'kick'] [ ]]
[X] 1018: EXIST_S_26 'there is' [REL_PAS_SVO_WHO_28 [ ] 'who is' [KICK_29 'kick'] '-ed by' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]]
[X] 1024: SVO_5 [REL_SVO_WHO_27 [WOMAN_11 'woman'] 'who' [KICK_29 'kick'] [ ]] [WEAR_10 'wear'] [ADJ_NOUN_17 [GREEN_19 'green'] [DRESS_12 'dress']]

> Next Attention
  RIGHT_WOMAN_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 5
================================================================================
> Current Attention
  RIGHT_WOMAN_AREA (perception done)

> Perceived Regions
  RIGHT_WOMAN_AREA

> Schema Instances
[ O] SemRep-N WOMAN_0
[ O] SemRep-N WEAR_1
[ O] SemRep-N DRESS_2
[ O] SemRep-R AGENT_3 from WEAR_1 to WOMAN_0
[ O] SemRep-R PATIENT_4 from WEAR_1 to DRESS_2
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [REL_SVO_WHO_40]
[ O] Construction EXIST_S_8 covering DRESS_2 for 'there is' [ ]
[ O] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ O] SemRep-N GREEN_13
[ O] SemRep-R MODIFY_14 from GREEN_13 to DRESS_2
[ O] Construction IN_COLOR_18 covering WOMAN_0 WEAR_1 DRESS_2 GREEN_13 AGENT_3 PATIENT_4 MODIFY_14 for [WOMAN_11] 'in' [GREEN_19]
[ O] Construction GREEN_19 covering GREEN_13 for 'green'
[ O] SemRep-N KICK_20
[ O] SemRep-R PATIENT_21 from KICK_20 to WOMAN_23
[ O] SemRep-R AGENT_22 from KICK_20 to WOMAN_0
[!O] SemRep-N WOMAN_23
[ O] Construction SVO_24 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [IN_COLOR_18] [KICK_29] [REL_SVO_WHO_41]
[ O] Construction EXIST_S_26 covering WOMAN_23 for 'there is' [REL_PAS_SVO_WHO_42]
[ O] Construction KICK_29 covering KICK_20 for 'kick'
[!O] SemRep-N WEAR_30
[!O] SemRep-N DRESS_31
[!O] SemRep-R AGENT_32 from WEAR_30 to WOMAN_23
[!O] SemRep-R PATIENT_33 from WEAR_30 to DRESS_31
[!X] Construction SVO_35 covering WOMAN_23 DRESS_31 WEAR_30 AGENT_32 PATIENT_33 for [REL_PAS_SVO_WHO_42] [WEAR_43] [DRESS_45]
[!X] Construction PAS_SVO_36 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [REL_SVO_WHO_41] 'is' [KICK_29] '-ed by' [IN_COLOR_18]
[!X] Construction PAS_SVO_37 covering WOMAN_23 DRESS_31 WEAR_30 AGENT_32 PATIENT_33 for [DRESS_45] 'is' [WEAR_43] '-ed by' [REL_PAS_SVO_WHO_42]
[!O] Construction EXIST_S_39 covering DRESS_31 for 'there is' [DRESS_45]
[!X] Construction REL_SVO_WHO_40 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [IN_COLOR_18] 'who' [KICK_29] [REL_SVO_WHO_41]
[!O] Construction REL_SVO_WHO_41 covering WOMAN_23 DRESS_31 WEAR_30 AGENT_32 PATIENT_33 for [WOMAN_44] 'who' [WEAR_43] [DRESS_45]
[!X] Construction REL_PAS_SVO_WHO_42 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [WOMAN_44] 'who is' [KICK_29] '-ed by' [IN_COLOR_18]
[!O] Construction WEAR_43 covering WEAR_30 for 'wear'
```

```
[!O] Construction WOMAN_44 covering WOMAN_23 for 'woman'
[!O] Construction DRESS_45 covering DRESS_31 for 'dress'

> Competition Traces
  SVO_24(1517) eliminated PAS_SVO_36(1461)
  SVO_24(1517) eliminated REL_SVO_WHO_40(1507)
  SVO_24(1517) eliminated REL_PAS_SVO_WHO_42(1511)
  SVO_35(1511) eliminated PAS_SVO_37(1455)
  REL_SVO_WHO_41(1517) eliminated SVO_35(1511)

> Construction Structures
[ ] 43: EXIST_S_8 'there is' [ ]
[ ] 138: EXIST_S_7 'there is' [WOMAN_11 'woman']
[ ] 138: EXIST_S_26 'there is' [WOMAN_44 'woman']
[ ] 138: EXIST_S_39 'there is' [DRESS_45 'dress']
[ ] 731: EXIST_S_7 'there is' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]
[ ] 536: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [WOMAN_44 'woman']
[X] 480: PAS_SVO_36 [WOMAN_44 'woman'] 'is' [KICK_29 'kick'] '-ed by' [WOMAN_11 'woman']
[X] 536: SVO_35 [WOMAN_44 'woman'] [WEAR_43 'wear'] [DRESS_45 'dress']
[X] 480: PAS_SVO_37 [DRESS_45 'dress'] 'is' [WEAR_43 'wear'] '-ed by' [WOMAN_44 'woman']
[ ] 1129: SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [WOMAN_44 'woman']
[X] 1073: PAS_SVO_36 [WOMAN_44 'woman'] 'is' [KICK_29 'kick'] '-ed by' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]
[X] 526: EXIST_S_7 'there is' [REL_SVO_WHO_40 [WOMAN_11 'woman'] 'who' [KICK_29 'kick'] [WOMAN_44 'woman']]
[X] 1119: EXIST_S_7 'there is' [REL_SVO_WHO_40 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] 'who' [KICK_29 'kick'] [WOMAN_44 'woman']]
[X] 520: EXIST_S_26 'there is' [REL_PAS_SVO_WHO_42 [WOMAN_44 'woman'] 'who is' [KICK_29 'kick'] '-ed by' [WOMAN_11 'woman']]
[ ] 526: EXIST_S_26 'there is' [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [DRESS_45 'dress']]
[X] 1113: EXIST_S_26 'there is' [REL_PAS_SVO_WHO_42 [WOMAN_44 'woman'] 'who is' [KICK_29 'kick'] '-ed by' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19
'green']]]
[ ] 924: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [DRESS_45 'dress']]
[X] 868: PAS_SVO_36 [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [DRESS_45 'dress']] 'is' [KICK_29 'kick'] '-ed by' [WOMAN_11 'woman']
[X] 918: SVO_35 [REL_PAS_SVO_WHO_42 [WOMAN_44 'woman'] 'who is' [KICK_29 'kick'] '-ed by' [WOMAN_11 'woman']] [WEAR_43 'wear'] [DRESS_45 'dress']
[X] 1511: SVO_35 [REL_PAS_SVO_WHO_42 [WOMAN_44 'woman'] 'who is' [KICK_29 'kick'] '-ed by' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]]
[WEAR_43 'wear'] [DRESS_45 'dress']
[X] 862: PAS_SVO_37 [DRESS_45 'dress'] 'is' [WEAR_43 'wear'] '-ed by' [REL_PAS_SVO_WHO_42 [WOMAN_44 'woman'] 'who is' [KICK_29 'kick'] '-ed by'
[WOMAN_11 'woman']]
[X] 1455: PAS_SVO_37 [DRESS_45 'dress'] 'is' [WEAR_43 'wear'] '-ed by' [REL_PAS_SVO_WHO_42 [WOMAN_44 'woman'] 'who is' [KICK_29 'kick'] '-ed by'
[IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]]
[ ] 1517: SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear']
[DRESS_45 'dress']]
[X] 1461: PAS_SVO_36 [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [DRESS_45 'dress']] 'is' [KICK_29 'kick'] '-ed by' [IN_COLOR_18 [WOMAN_11
'woman'] 'in' [GREEN_19 'green']]
[X] 914: EXIST_S_7 'there is' [REL_SVO_WHO_40 [WOMAN_11 'woman'] 'who' [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear']
[DRESS_45 'dress']]]
[X] 1507: EXIST_S_7 'there is' [REL_SVO_WHO_40 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] 'who' [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44
'woman'] 'who' [WEAR_43 'wear'] [DRESS_45 'dress']]]

> Next Attention
  RIGHT_DRESS_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 6
================================================================================
> Current Attention
  RIGHT_DRESS_AREA (perception done)

> Perceived Regions
  RIGHT_DRESS_AREA

> Schema Instances
[ O] SemRep-N WOMAN_0
[ O] SemRep-N WEAR_1
[ O] SemRep-N DRESS_2
[ O] SemRep-R AGENT_3 from WEAR_1 to WOMAN_0
[ O] SemRep-R PATIENT_4 from WEAR_1 to DRESS_2
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [IN_COLOR_18]
[ O] Construction EXIST_S_8 covering DRESS_2 for 'there is' [ ]
[ O] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ O] SemRep-N GREEN_13
[ O] SemRep-R MODIFY_14 from GREEN_13 to DRESS_2
[ O] Construction IN_COLOR_18 covering WOMAN_0 WEAR_1 DRESS_2 GREEN_13 AGENT_3 PATIENT_4 MODIFY_14 for [WOMAN_11] 'in' [GREEN_19]
[ O] Construction GREEN_19 covering GREEN_13 for 'green'
[ O] SemRep-N KICK_20
[ O] SemRep-R PATIENT_21 from KICK_20 to WOMAN_23
[ O] SemRep-R AGENT_22 from KICK_20 to WOMAN_0
[ O] SemRep-N WOMAN_23
[ O] Construction SVO_24 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [IN_COLOR_18] [KICK_29] [IN_COLOR_51]
[ O] Construction EXIST_S_26 covering WOMAN_23 for 'there is' [IN_COLOR_51]
[ O] Construction KICK_29 covering KICK_20 for 'kick'
[ O] SemRep-N WEAR_30
[ O] SemRep-N DRESS_31
[ O] SemRep-R AGENT_32 from WEAR_30 to WOMAN_23
[ O] SemRep-R PATIENT_33 from WEAR_30 to DRESS_31
[ O] Construction EXIST_S_39 covering DRESS_31 for 'there is' [ADJ_NOUN_50]
[ X] Construction REL_SVO_WHO_41 covering WOMAN_23 DRESS_31 WEAR_30 AGENT_32 PATIENT_33 for [WOMAN_44] 'who' [WEAR_43] [ADJ_NOUN_50]
[ X] Construction WEAR_43 covering WEAR_30 for 'wear'
[ O] Construction WOMAN_44 covering WOMAN_23 for 'woman'
[ X] Construction DRESS_45 covering DRESS_31 for 'dress'
[!O] SemRep-N BLUE_46
[!O] SemRep-R MODIFY_47 from BLUE_46 to DRESS_31
[!X] Construction SPA_48 covering DRESS_31 BLUE_46 MODIFY_47 for [DRESS_45] 'is' [BLUE_52]
[!X] Construction REL_SPA_WHICH_49 covering DRESS_31 BLUE_46 MODIFY_47 for [DRESS_45] 'which is' [BLUE_52]
[!X] Construction ADJ_NOUN_50 covering DRESS_31 BLUE_46 MODIFY_47 for [BLUE_52] [DRESS_45]
[!O] Construction IN_COLOR_51 covering WOMAN_23 WEAR_30 DRESS_31 BLUE_46 AGENT_32 PATIENT_33 MODIFY_47 for [WOMAN_44] 'in' [BLUE_52]
[!O] Construction BLUE_52 covering BLUE_46 for 'blue'

> Competition Traces
  IN_COLOR_51(1723) eliminated REL_SVO_WHO_41(1713)
  IN_COLOR_51(1723) eliminated WEAR_43(1713)
  IN_COLOR_51(1723) eliminated DRESS_45(1713)
  REL_SPA_WHICH_49(1706) eliminated SPA_48(339)
  ADJ_NOUN_50(1713) eliminated REL_SPA_WHICH_49(1706)
  IN_COLOR_51(1723) eliminated ADJ_NOUN_50(1713)

> Construction Structures
[ ] 43: EXIST_S_8 'there is' [ ]
[ ] 138: EXIST_S_7 'there is' [WOMAN_11 'woman']
[ ] 138: EXIST_S_26 'there is' [WOMAN_44 'woman']
```

```
[X] 138: EXIST_S_39 'there is' [DRESS_45 'dress']
[X] 339: SPA_48 [DRESS_45 'dress'] 'is' [BLUE_52 'blue']
[ ] 731: EXIST_S_7 'there is' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]
[ ] 536: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [WOMAN_44 'woman']
[ ] 732: EXIST_S_26 'there is' [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]
[X] 327: EXIST_S_39 'there is' [REL_SPA_WHICH_49 [DRESS_45 'dress'] 'which is' [BLUE_52 'blue']]
[X] 334: EXIST_S_39 'there is' [ADJ_NOUN_50 [BLUE_52 'blue'] [DRESS_45 'dress']]
[ ] 1129: SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [WOMAN_44 'woman']
[ ] 1130: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]
[X] 526: EXIST_S_26 'there is' [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [DRESS_45 'dress']]
[X] 715: EXIST_S_26 'there is' [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [REL_SPA_WHICH_49 [DRESS_45 'dress'] 'which is' [BLUE_52
'blue']]]
[X] 722: EXIST_S_26 'there is' [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [ADJ_NOUN_50 [BLUE_52 'blue'] [DRESS_45 'dress']]]
[ ] 1723: SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]
[X] 924: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [DRESS_45 'dress']]
[X] 1113: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [REL_SPA_WHICH_49 [DRESS_45 'dress']
'which is' [BLUE_52 'blue']]]
[X] 1120: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [ADJ_NOUN_50 [BLUE_52 'blue'] [DRESS_45
'dress']]]
[X] 1517: SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear']
[DRESS_45 'dress']]
[X] 1706: SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear']
[REL_SPA_WHICH_49 [DRESS_45 'dress'] 'which is' [BLUE_52 'blue']]]
[X] 1713: SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear']
[ADJ_NOUN_50 [BLUE_52 'blue'] [DRESS_45 'dress']]]

> Next Attention
  BOXINGRING_AREA (uncertainty left: 2)

================================================================================
  Simulation Time: 7
================================================================================
> Current Attention
  BOXINGRING_AREA (uncertainty left: 1)

> Schema Instances
[ O] SemRep-N WOMAN_0
[ O] SemRep-N WEAR_1
[ O] SemRep-N DRESS_2
[ O] SemRep-R AGENT_3 from WEAR_1 to WOMAN_0
[ O] SemRep-R PATIENT_4 from WEAR_1 to DRESS_2
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [IN_COLOR_18]
[ O] Construction EXIST_S_8 covering DRESS_2 for 'there is' [ ]
[ O] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ O] SemRep-N GREEN_13
[ O] SemRep-R MODIFY_14 from GREEN_13 to DRESS_2
[ O] Construction IN_COLOR_18 covering WOMAN_0 WEAR_1 DRESS_2 GREEN_13 AGENT_3 PATIENT_4 MODIFY_14 for [WOMAN_11] 'in' [GREEN_19]
[ O] Construction GREEN_19 covering GREEN_13 for 'green'
[ O] SemRep-N KICK_20
[ O] SemRep-R PATIENT_21 from KICK_20 to WOMAN_23
[ O] SemRep-R AGENT_22 from KICK_20 to WOMAN_0
[ O] SemRep-N WOMAN_23
[ O] Construction SVO_24 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [IN_COLOR_18] [KICK_29] [IN_COLOR_51]
[ O] Construction EXIST_S_26 covering WOMAN_23 for 'there is' [IN_COLOR_51]
[ O] Construction KICK_29 covering KICK_20 for 'kick'
[ O] SemRep-N WEAR_30
[ O] SemRep-N DRESS_31
[ O] SemRep-R AGENT_32 from WEAR_30 to WOMAN_23
[ O] SemRep-R PATIENT_33 from WEAR_30 to DRESS_31
[ O] Construction EXIST_S_39 covering DRESS_31 for 'there is' [ ]
[ O] Construction WOMAN_44 covering WOMAN_23 for 'woman'
[ O] SemRep-N BLUE_46
[ O] SemRep-R MODIFY_47 from BLUE_46 to DRESS_31
[ O] Construction IN_COLOR_51 covering WOMAN_23 WEAR_30 DRESS_31 BLUE_46 AGENT_32 PATIENT_33 MODIFY_47 for [WOMAN_44] 'in' [BLUE_52]
[ O] Construction BLUE_52 covering BLUE_46 for 'blue'

> Construction Structures
[ ] 43: EXIST_S_8 'there is' [ ]
[ ] 43: EXIST_S_39 'there is' [ ]
[ ] 138: EXIST_S_7 'there is' [WOMAN_11 'woman']
[ ] 138: EXIST_S_26 'there is' [WOMAN_44 'woman']
[ ] 731: EXIST_S_7 'there is' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]
[ ] 536: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [WOMAN_44 'woman']
[ ] 732: EXIST_S_26 'there is' [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]
[ ] 1129: SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [WOMAN_44 'woman']
[ ] 1130: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]
[ ] 1723: SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]

> Next Attention
  BOXINGRING_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 8
================================================================================
> Current Attention
  BOXINGRING_AREA (perception done)

> Perceived Regions
  BOXINGRING_AREA

> Schema Instances
[ O] SemRep-N WOMAN_0
[ O] SemRep-N WEAR_1
[ O] SemRep-N DRESS_2
[ O] SemRep-R AGENT_3 from WEAR_1 to WOMAN_0
[ O] SemRep-R PATIENT_4 from WEAR_1 to DRESS_2
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [IN_COLOR_18]
[ O] Construction EXIST_S_8 covering DRESS_2 for 'there is' [ ]
[ O] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ O] SemRep-N GREEN_13
[ O] SemRep-R MODIFY_14 from GREEN_13 to DRESS_2
[ O] Construction IN_COLOR_18 covering WOMAN_0 WEAR_1 DRESS_2 GREEN_13 AGENT_3 PATIENT_4 MODIFY_14 for [WOMAN_11] 'in' [GREEN_19]
[ O] Construction GREEN_19 covering GREEN_13 for 'green'
[ O] SemRep-N KICK_20
[ O] SemRep-R PATIENT_21 from KICK_20 to WOMAN_23
[ O] SemRep-R AGENT_22 from KICK_20 to WOMAN_0
[ O] SemRep-N WOMAN_23
```

```
[ O] Construction SVO_24 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [IN_COLOR_18] [KICK_29] [IN_COLOR_51]
[ O] Construction EXIST_S_26 covering WOMAN_23 for 'there is' [IN_COLOR_51]
[ O] Construction KICK_29 covering KICK_20 for 'kick'
[ O] SemRep-N WEAR_30
[ O] SemRep-N DRESS_31
[ O] SemRep-R AGENT_32 from WEAR_30 to WOMAN_23
[ O] SemRep-R PATIENT_33 from WEAR_30 to DRESS_31
[ O] Construction EXIST_S_39 covering DRESS_31 for 'there is' [ ]
[ O] Construction WOMAN_44 covering WOMAN_23 for 'woman'
[ O] SemRep-N BLUE_46
[ O] SemRep-R MODIFY_47 from BLUE_46 to DRESS_31
[ O] Construction IN_COLOR_51 covering WOMAN_23 WEAR_30 DRESS_31 BLUE_46 AGENT_32 PATIENT_33 MODIFY_47 for [WOMAN_44] 'in' [BLUE_52]
[ O] Construction BLUE_52 covering BLUE_46 for 'blue'
[!O] SemRep-N BOXINGRING_53
[!O] SemRep-R IN_54 from KICK_20 to BOXINGRING_53
[!O] Construction THEME_S_55 covering BOXINGRING_53 for 'it is' [BOXINGRING_57]
[!O] Construction PP_IN_56 covering KICK_20 IN_54 BOXINGRING_53 for [SVO_24] 'in' [BOXINGRING_57]
[!O] Construction BOXINGRING_57 covering BOXINGRING_53 for 'boxing ring'

> Construction Structures
[ ] 43: EXIST_S_8 'there is' [ ]
[ ] 43: EXIST_S_39 'there is' [ ]
[ ] 138: EXIST_S_7 'there is' [WOMAN_11 'woman']
[ ] 138: EXIST_S_26 'there is' [WOMAN_44 'woman']
[ ] 136: THEME_S_55 'it is' [BOXINGRING_57 'boxing ring']
[ ] 731: EXIST_S_7 'there is' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]
[ ] 732: EXIST_S_26 'there is' [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]
[ ] 774: PP_IN_56 [SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [WOMAN_44 'woman']] 'in' [BOXINGRING_57 'boxing ring']
[ ] 1367: PP_IN_56 [SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [WOMAN_44 'woman']] 'in' [BOXINGRING_57 'boxing
ring']
[ ] 1368: PP_IN_56 [SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]] 'in' [BOXINGRING_57 'boxing
ring']
[ ] 1961: PP_IN_56 [SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52
'blue']]] 'in' [BOXINGRING_57 'boxing ring']

> Next Attention
  PEOPLE_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 9
================================================================================
> Current Attention
  PEOPLE_AREA (perception done)

> Perceived Regions
  PEOPLE_AREA

> Schema Instances
[ O] SemRep-N WOMAN_0
[ O] SemRep-N WEAR_1
[ O] SemRep-N DRESS_2
[ O] SemRep-R AGENT_3 from WEAR_1 to WOMAN_0
[ O] SemRep-R PATIENT_4 from WEAR_1 to DRESS_2
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [IN_COLOR_18]
[ O] Construction EXIST_S_8 covering DRESS_2 for 'there is' [ ]
[ O] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ O] SemRep-N GREEN_13
[ O] SemRep-R MODIFY_14 from GREEN_13 to DRESS_2
[ O] Construction IN_COLOR_18 covering WOMAN_0 WEAR_1 DRESS_2 GREEN_13 AGENT_3 PATIENT_4 MODIFY_14 for [WOMAN_11] 'in' [GREEN_19]
[ O] Construction GREEN_19 covering GREEN_13 for 'green'
[ O] SemRep-N KICK_20
[ O] SemRep-R PATIENT_21 from KICK_20 to WOMAN_23
[ O] SemRep-R AGENT_22 from KICK_20 to WOMAN_0
[ O] SemRep-N WOMAN_23
[ O] Construction SVO_24 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [IN_COLOR_18] [KICK_29] [IN_COLOR_51]
[ O] Construction EXIST_S_26 covering WOMAN_23 for 'there is' [IN_COLOR_51]
[ O] Construction KICK_29 covering KICK_20 for 'kick'
[ O] SemRep-N WEAR_30
[ O] SemRep-N DRESS_31
[ O] SemRep-R AGENT_32 from WEAR_30 to WOMAN_23
[ O] SemRep-R PATIENT_33 from WEAR_30 to DRESS_31
[ O] Construction EXIST_S_39 covering DRESS_31 for 'there is' [ ]
[ O] Construction WOMAN_44 covering WOMAN_23 for 'woman'
[ O] SemRep-N BLUE_46
[ O] SemRep-R MODIFY_47 from BLUE_46 to DRESS_31
[ O] Construction IN_COLOR_51 covering WOMAN_23 WEAR_30 DRESS_31 BLUE_46 AGENT_32 PATIENT_33 MODIFY_47 for [WOMAN_44] 'in' [BLUE_52]
[ O] Construction BLUE_52 covering BLUE_46 for 'blue'
[ O] SemRep-N BOXINGRING_53
[ O] SemRep-R IN_54 from KICK_20 to BOXINGRING_53
[ O] Construction THEME_S_55 covering BOXINGRING_53 for 'it is' [BOXINGRING_57]
[ O] Construction PP_IN_56 covering KICK_20 IN_54 BOXINGRING_53 for [SVO_24] 'in' [BOXINGRING_57]
[ O] Construction BOXINGRING_57 covering BOXINGRING_53 for 'boxing ring'
[!O] SemRep-N PEOPLE_58
[!O] Construction EXIST_S_59 covering PEOPLE_58 for 'there is' [PEOPLE_60]
[!O] Construction PEOPLE_60 covering PEOPLE_58 for 'people'

> Construction Structures
[ ] 43: EXIST_S_8 'there is' [ ]
[ ] 43: EXIST_S_39 'there is' [ ]
[ ] 138: EXIST_S_7 'there is' [WOMAN_11 'woman']
[ ] 138: EXIST_S_26 'there is' [WOMAN_44 'woman']
[ ] 136: THEME_S_55 'it is' [BOXINGRING_57 'boxing ring']
[ ] 137: EXIST_S_59 'there is' [PEOPLE_60 'people']
[ ] 731: EXIST_S_7 'there is' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]
[ ] 732: EXIST_S_26 'there is' [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]
[ ] 774: PP_IN_56 [SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [WOMAN_44 'woman']] 'in' [BOXINGRING_57 'boxing ring']
[ ] 1367: PP_IN_56 [SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [WOMAN_44 'woman']] 'in' [BOXINGRING_57 'boxing
ring']
[ ] 1368: PP_IN_56 [SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]] 'in' [BOXINGRING_57 'boxing
ring']
[ ] 1961: PP_IN_56 [SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52
'blue']]] 'in' [BOXINGRING_57 'boxing ring']

> Next Attention
  PEOPLE_FOCUS_AREA (uncertainty left: 1)

================================================================================
```

```
  Simulation Time: 10
================================================================================
> Current Attention
  PEOPLE_FOCUS_AREA (perception done)

> Perceived Regions
  PEOPLE_FOCUS_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ @] SemRep-N WEAR_1
[ @] SemRep-N DRESS_2
[ @] SemRep-R AGENT_3 from WEAR_1 to WOMAN_0
[ @] SemRep-R PATIENT_4 from WEAR_1 to DRESS_2
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [IN_COLOR_18]
[ O] Construction EXIST_S_8 covering DRESS_2 for 'there is' [ ]
[ @] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ @] SemRep-N GREEN_13
[ @] SemRep-R MODIFY_14 from GREEN_13 to DRESS_2
[ @] Construction IN_COLOR_18 covering WOMAN_0 WEAR_1 DRESS_2 GREEN_13 AGENT_3 PATIENT_4 MODIFY_14 for [WOMAN_11] 'in' [GREEN_19]
[ @] Construction GREEN_19 covering GREEN_13 for 'green'
[ @] SemRep-N KICK_20
[ @] SemRep-R PATIENT_21 from KICK_20 to WOMAN_23
[ @] SemRep-R AGENT_22 from KICK_20 to WOMAN_0
[ @] SemRep-N WOMAN_23
[ @] Construction SVO_24 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [IN_COLOR_18] [KICK_29] [IN_COLOR_51]
[ O] Construction EXIST_S_26 covering WOMAN_23 for 'there is' [IN_COLOR_51]
[ @] Construction KICK_29 covering KICK_20 for 'kick'
[ @] SemRep-N WEAR_30
[ @] SemRep-N DRESS_31
[ @] SemRep-R AGENT_32 from WEAR_30 to WOMAN_23
[ @] SemRep-R PATIENT_33 from WEAR_30 to DRESS_31
[ O] Construction EXIST_S_39 covering DRESS_31 for 'there is' [ ]
[ @] Construction WOMAN_44 covering WOMAN_23 for 'woman'
[ @] SemRep-N BLUE_46
[ @] SemRep-R MODIFY_47 from BLUE_46 to DRESS_31
[ @] Construction IN_COLOR_51 covering WOMAN_23 WEAR_30 DRESS_31 BLUE_46 AGENT_32 PATIENT_33 MODIFY_47 for [WOMAN_44] 'in' [BLUE_52]
[ @] Construction BLUE_52 covering BLUE_46 for 'blue'
[ @] SemRep-N BOXINGRING_53
[ @] SemRep-R IN_54 from KICK_20 to BOXINGRING_53
[ O] Construction THEME_S_55 covering BOXINGRING_53 for 'it is' [BOXINGRING_57]
[ @] Construction PP_IN_56 covering KICK_20 IN_54 BOXINGRING_53 for [SVO_24] 'in' [BOXINGRING_57]
[ @] Construction BOXINGRING_57 covering BOXINGRING_53 for 'boxing ring'
[ @] SemRep-N PEOPLE_58
[ O] Construction EXIST_S_59 covering PEOPLE_58 for 'there is' [REL_SV_WHO_66]
[ @] Construction PEOPLE_60 covering PEOPLE_58 for 'people'
[!@] SemRep-N WATCH_61
[!@] SemRep-R AGENT_62 from WATCH_61 to PEOPLE_58
[!@] SemRep-R CONCURRENT_63 from KICK_20 to WATCH_61
[!@] Construction CNJ_WHILE_64 covering KICK_20 WATCH_61 CONCURRENT_63 for [PP_IN_56] 'while' [SV_65]
[!@] Construction SV_65 covering PEOPLE_58 WATCH_61 AGENT_62 for [PEOPLE_60] [WATCH_67]
[!X] Construction REL_SV_WHO_66 covering PEOPLE_58 WATCH_61 AGENT_62 for [PEOPLE_60] 'who' [WATCH_67]
[!@] Construction WATCH_67 covering WATCH_61 for 'watch'

> Competition Traces
  SV_65(2395) eliminated REL_SV_WHO_66(329)

> Construction Structures
[ ] 43: EXIST_S_8 'there is' [ ]
[ ] 43: EXIST_S_39 'there is' [ ]
[ ] 138: EXIST_S_7 'there is' [WOMAN_11 'woman']
[ ] 138: EXIST_S_26 'there is' [WOMAN_44 'woman']
[ ] 136: THEME_S_55 'it is' [BOXINGRING_57 'boxing ring']
[ ] 137: EXIST_S_59 'there is' [PEOPLE_60 'people']
[ ] 731: EXIST_S_7 'there is' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]
[ ] 732: EXIST_S_26 'there is' [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]
[X] 329: EXIST_S_59 'there is' [REL_SV_WHO_66 [PEOPLE_60 'people'] 'who' [WATCH_67 'watch']]
[ ] 970: CNJ_WHILE_64 [SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [WOMAN_44 'woman']] 'while' [SV_65 [PEOPLE_60 'people'] [WATCH_67 'watch']]
[ ] 1563: CNJ_WHILE_64 [SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [WOMAN_44 'woman']] 'while' [SV_65 [PEOPLE_60
'people'] [WATCH_67 'watch']]
[ ] 1564: CNJ_WHILE_64 [SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]] 'while' [SV_65 [PEOPLE_60
'people'] [WATCH_67 'watch']]
[ ] 1208: CNJ_WHILE_64 [PP_IN_56 [SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [WOMAN_44 'woman']] 'in' [BOXINGRING_57 'boxing ring']] 'while' [SV_65
[PEOPLE_60 'people'] [WATCH_67 'watch']]
[ ] 1801: CNJ_WHILE_64 [PP_IN_56 [SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [WOMAN_44 'woman']] 'in'
[BOXINGRING_57 'boxing ring']] 'while' [SV_65 [PEOPLE_60 'people'] [WATCH_67 'watch']]
[ ] 1802: CNJ_WHILE_64 [PP_IN_56 [SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]] 'in' [BOXINGRING_57
'boxing ring']] 'while' [SV_65 [PEOPLE_60 'people'] [WATCH_67 'watch']]
[ ] 2157: CNJ_WHILE_64 [SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52
'blue']]] 'while' [SV_65 [PEOPLE_60 'people'] [WATCH_67 'watch']]
[*] 2395: CNJ_WHILE_64 [PP_IN_56 [SVO_24 [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']] [KICK_29 'kick'] [IN_COLOR_51 [WOMAN_44 'woman'] 'in'
[BLUE_52 'blue']]] 'in' [BOXINGRING_57 'boxing ring']] 'while' [SV_65 [PEOPLE_60 'people'] [WATCH_67 'watch']]

> Produced Utterance
  "woman in green kick woman in blue in boxing ring while people watch"

> Next Attention
  None

================================================================================
  Simulation Time: 11
================================================================================
> Current Attention
  None

> Schema Instances
[ x] SemRep-N WOMAN_0
[ x] SemRep-N WEAR_1
[ x] SemRep-N DRESS_2
[ x] SemRep-R AGENT_3 from WEAR_1 to WOMAN_0
[ x] SemRep-R PATIENT_4 from WEAR_1 to DRESS_2
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [ ]
[ O] Construction EXIST_S_8 covering DRESS_2 for 'there is' [ ]
[ x] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ x] SemRep-N GREEN_13
[ x] SemRep-R MODIFY_14 from GREEN_13 to DRESS_2
[ x] Construction IN_COLOR_18 covering WOMAN_0 WEAR_1 DRESS_2 GREEN_13 AGENT_3 PATIENT_4 MODIFY_14 for [ ] 'in' [ ]
```

```
[ x] Construction GREEN_19 covering GREEN_13 for 'green'
[ x] SemRep-N KICK_20
[ x] SemRep-R PATIENT_21 from KICK_20 to WOMAN_23
[ x] SemRep-R AGENT_22 from KICK_20 to WOMAN_0
[ x] SemRep-N WOMAN_23
[ x] Construction SVO_24 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [ ] [ ] [ ]
[ O] Construction EXIST_S_26 covering WOMAN_23 for 'there is' [ ]
[ x] Construction KICK_29 covering KICK_20 for 'kick'
[ x] SemRep-N WEAR_30
[ x] SemRep-N DRESS_31
[ x] SemRep-R AGENT_32 from WEAR_30 to WOMAN_23
[ x] SemRep-R PATIENT_33 from WEAR_30 to DRESS_31
[ O] Construction EXIST_S_39 covering DRESS_31 for 'there is' [ ]
[ x] Construction WOMAN_44 covering WOMAN_23 for 'woman'
[ x] SemRep-N BLUE_46
[ x] SemRep-R MODIFY_47 from BLUE_46 to DRESS_31
[ x] Construction IN_COLOR_51 covering WOMAN_23 WEAR_30 DRESS_31 BLUE_46 AGENT_32 PATIENT_33 MODIFY_47 for [ ] 'in' [ ]
[ x] Construction BLUE_52 covering BLUE_46 for 'blue'
[ x] SemRep-N BOXINGRING_53
[ x] SemRep-R IN_54 from KICK_20 to BOXINGRING_53
[ O] Construction THEME_S_55 covering BOXINGRING_53 for 'it is' [ ]
[ x] Construction PP_IN_56 covering KICK_20 IN_54 BOXINGRING_53 for [ ] 'in' [ ]
[ x] Construction BOXINGRING_57 covering BOXINGRING_53 for 'boxing ring'
[ x] SemRep-N PEOPLE_58
[ O] Construction EXIST_S_59 covering PEOPLE_58 for 'there is' [ ]
[ x] Construction PEOPLE_60 covering PEOPLE_58 for 'people'
[ x] SemRep-N WATCH_61
[ x] SemRep-R AGENT_62 from WATCH_61 to PEOPLE_58
[ x] SemRep-R CONCURRENT_63 from KICK_20 to WATCH_61
[ x] Construction CNJ_WHILE_64 covering KICK_20 WATCH_61 CONCURRENT_63 for [ ] 'while' [ ]
[ x] Construction SV_65 covering PEOPLE_58 WATCH_61 AGENT_62 for [ ] [ ]
[ x] Construction WATCH_67 covering WATCH_61 for 'watch'

> Next Attention
  None

================================================================================
  Simulation Time: 12
================================================================================
> Current Attention
  None

> Next Attention
  None


Simulation complete: inactivity termination.
```

The following is the simulation output for the low threshold case. Only the time parameter is tuned to “1” while the others are left to be infinite in order to simulate the time pressure induced threshold effect.

```
Template Construction Grammar (TCG) Simulator v2.5

Jinyong Lee (jinyongl@usc.edu), June 23. 2012
USC Brain Project, Computer Science Department
University of Southern California (USC)

Loading Initialization File 'TCG.ini'...
Loading Semantic Network 'TCG_semantics.txt'...
Loading Construction Vocabulary 'TCG_vocabulary.txt'...
Loading Scene 'scene_cholita.txt'...

Initializing Simulator...
- Max Simulation Time: 20
- Premature Production: on
- Utterance Continuity: on
- Verbal Guidance: on
- Threshold of Utterance: Time = 1, CNXs = infinite, Syllables = infinite

Beginning Simulation...

================================================================================
  Simulation Time: 1
================================================================================
> Current Attention
  None

> Next Attention
  LEFT_WOMAN_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 2
================================================================================
> Current Attention
  LEFT_WOMAN_AREA (perception done)

> Perceived Regions
  LEFT_WOMAN_AREA

> Schema Instances
[!@] SemRep-N WOMAN_0
[!@] SemRep-N WEAR_1
[!@] SemRep-N DRESS_2
[!@] SemRep-R AGENT_3 from WEAR_1 to WOMAN_0
[!@] SemRep-R PATIENT_4 from WEAR_1 to DRESS_2
[!@] Construction SVO_5 covering WOMAN_0 DRESS_2 WEAR_1 AGENT_3 PATIENT_4 for [WOMAN_11] [WEAR_10] [DRESS_12]
[!X] Construction PAS_SVO_6 covering WOMAN_0 DRESS_2 WEAR_1 AGENT_3 PATIENT_4 for [DRESS_12] 'is' [WEAR_10] '-ed by' [WOMAN_11]
[!O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [REL_SVO_WHO_9]
[!O] Construction EXIST_S_8 covering DRESS_2 for 'there is' [DRESS_12]
[!X] Construction REL_SVO_WHO_9 covering WOMAN_0 DRESS_2 WEAR_1 AGENT_3 PATIENT_4 for [WOMAN_11] 'who' [WEAR_10] [DRESS_12]
[!@] Construction WEAR_10 covering WEAR_1 for 'wear'
[!@] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[!@] Construction DRESS_12 covering DRESS_2 for 'dress'
```

```
> Competition Traces
  SVO_5(536) eliminated PAS_SVO_6(480)
  SVO_5(536) eliminated REL_SVO_WHO_9(526)

> Construction Structures
[ ] 138: EXIST_S_7 'there is' [WOMAN_11 'woman']
[ ] 138: EXIST_S_8 'there is' [DRESS_12 'dress']
[*] 536: SVO_5 [WOMAN_11 'woman'] [WEAR_10 'wear'] [DRESS_12 'dress']
[X] 480: PAS_SVO_6 [DRESS_12 'dress'] 'is' [WEAR_10 'wear'] '-ed by' [WOMAN_11 'woman']
[X] 526: EXIST_S_7 'there is' [REL_SVO_WHO_9 [WOMAN_11 'woman'] 'who' [WEAR_10 'wear'] [DRESS_12 'dress']]

> Produced Utterance
  "woman wear dress"

> Next Attention
  LEFT_DRESS_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 3
================================================================================
> Current Attention
  LEFT_DRESS_AREA (perception done)

> Perceived Regions
  LEFT_DRESS_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ @] SemRep-N WEAR_1
[ @] SemRep-N DRESS_2
[ @] SemRep-R AGENT_3 from WEAR_1 to WOMAN_0
[ @] SemRep-R PATIENT_4 from WEAR_1 to DRESS_2
[ @] Construction SVO_5 covering WOMAN_0 DRESS_2 WEAR_1 AGENT_3 PATIENT_4 for [WOMAN_11] [WEAR_10] [REL_SPA_WHICH_16]
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [IN_COLOR_18]
[ O] Construction EXIST_S_8 covering DRESS_2 for 'there is' [ADJ_NOUN_17]
[ @] Construction WEAR_10 covering WEAR_1 for 'wear'
[ @] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ @] Construction DRESS_12 covering DRESS_2 for 'dress'
[!@] SemRep-N GREEN_13
[!@] SemRep-R MODIFY_14 from GREEN_13 to DRESS_2
[!X] Construction SPA_15 covering DRESS_2 GREEN_13 MODIFY_14 for [DRESS_12] 'is' [GREEN_19]
[!@] Construction REL_SPA_WHICH_16 covering DRESS_2 GREEN_13 MODIFY_14 for [DRESS_12] 'which is' [GREEN_19]
[!X] Construction ADJ_NOUN_17 covering DRESS_2 GREEN_13 MODIFY_14 for [GREEN_19] [DRESS_12]
[!X] Construction IN_COLOR_18 covering WOMAN_0 WEAR_1 DRESS_2 GREEN_13 AGENT_3 PATIENT_4 MODIFY_14 for [WOMAN_11] 'in' [GREEN_19]
[!@] Construction GREEN_19 covering GREEN_13 for 'green'

> Competition Traces
  SVO_5(1324) eliminated IN_COLOR_18(631)
  REL_SPA_WHICH_16(1324) eliminated SPA_15(238)
  SPA_15(238) eliminated ADJ_NOUN_17(233)

> Construction Structures
[X] 238: SPA_15 [DRESS_12 'dress'] 'is' [GREEN_19 'green']
[X] 631: EXIST_S_7 'there is' [IN_COLOR_18 [WOMAN_11 'woman'] 'in' [GREEN_19 'green']]
[ ] 226: EXIST_S_8 'there is' [REL_SPA_WHICH_16 [DRESS_12 'dress'] 'which is' [GREEN_19 'green']]
[X] 233: EXIST_S_8 'there is' [ADJ_NOUN_17 [GREEN_19 'green'] [DRESS_12 'dress']]
[*] 1324: SVO_5 [WOMAN_11 'woman'] [WEAR_10 'wear'] [REL_SPA_WHICH_16 [DRESS_12 'dress'] 'which is' [GREEN_19 'green']]
[X] 131: SVO_5 [WOMAN_11 'woman'] [WEAR_10 'wear'] [ADJ_NOUN_17 [GREEN_19 'green'] [DRESS_12 'dress']]

> Produced Utterance
  "which is green"

> Next Attention
  KICK_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 4
================================================================================
> Current Attention
  KICK_AREA (perception done)

> Perceived Regions
  KICK_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ x] SemRep-N WEAR_1
[ x] SemRep-N DRESS_2
[ x] SemRep-R AGENT_3 from WEAR_1 to WOMAN_0
[ x] SemRep-R PATIENT_4 from WEAR_1 to DRESS_2
[ x] Construction SVO_5 covering WOMAN_0 DRESS_2 WEAR_1 AGENT_3 PATIENT_4 for [REL_SVO_WHO_27] [WEAR_10] [DRESS_12]
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [REL_SVO_WHO_27]
[ O] Construction EXIST_S_8 covering DRESS_2 for 'there is' [ ]
[ x] Construction WEAR_10 covering WEAR_1 for 'wear'
[ @] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ x] Construction DRESS_12 covering DRESS_2 for 'dress'
[ x] SemRep-N GREEN_13
[ x] SemRep-R MODIFY_14 from GREEN_13 to DRESS_2
[ x] Construction REL_SPA_WHICH_16 covering DRESS_2 GREEN_13 MODIFY_14 for [DRESS_12] 'which is' [GREEN_19]
[ x] Construction GREEN_19 covering GREEN_13 for 'green'
[!@] SemRep-N KICK_20
[!@] SemRep-R PATIENT_21 from KICK_20 to HUMAN_23
[!@] SemRep-R AGENT_22 from KICK_20 to WOMAN_0
[!O] SemRep-N HUMAN_23
[!@] Construction SVO_24 covering WOMAN_0 HUMAN_23 KICK_20 AGENT_22 PATIENT_21 for [WOMAN_11] [KICK_29] [ ]
[!X] Construction PAS_SVO_25 covering WOMAN_0 HUMAN_23 KICK_20 AGENT_22 PATIENT_21 for [ ] 'is' [KICK_29] '-ed by' [WOMAN_11]
[!O] Construction EXIST_S_26 covering HUMAN_23 for 'there is' [REL_PAS_SVO_WHO_28]
[!X] Construction REL_SVO_WHO_27 covering WOMAN_0 HUMAN_23 KICK_20 AGENT_22 PATIENT_21 for [WOMAN_11] 'who' [KICK_29] [ ]
[!X] Construction REL_PAS_SVO_WHO_28 covering WOMAN_0 HUMAN_23 KICK_20 AGENT_22 PATIENT_21 for [ ] 'who is' [KICK_29] '-ed by' [WOMAN_11]
[!@] Construction KICK_29 covering KICK_20 for 'kick'

> Competition Traces
  SVO_24(341) eliminated PAS_SVO_25(285)
  SVO_24(341) eliminated REL_SVO_WHO_27(331)
  SVO_24(341) eliminated REL_PAS_SVO_WHO_28(325)
```

```
> Construction Structures
[*] 341: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [ ]
[X] 285: PAS_SVO_25 [ ] 'is' [KICK_29 'kick'] '-ed by' [WOMAN_11 'woman']
[X] 331: EXIST_S_7 'there is' [REL_SVO_WHO_27 [WOMAN_11 'woman'] 'who' [KICK_29 'kick'] [ ]]
[X] 325: EXIST_S_26 'there is' [REL_PAS_SVO_WHO_28 [ ] 'who is' [KICK_29 'kick'] '-ed by' [WOMAN_11 'woman']]
[X] 329: SVO_5 [REL_SVO_WHO_27 [WOMAN_11 'woman'] 'who' [KICK_29 'kick'] [ ]] [WEAR_10 'wear'] [DRESS_12 'dress']
[X] 17: SVO_5 [REL_SVO_WHO_27 [WOMAN_11 'woman'] 'who' [KICK_29 'kick'] [ ]] [WEAR_10 'wear'] [REL_SPA_WHICH_16 [DRESS_12 'dress'] 'which is' [GREEN_19
'green']]

> Produced Utterance
  "woman kick..."

> Next Attention
  RIGHT_WOMAN_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 5
================================================================================
> Current Attention
  RIGHT_WOMAN_AREA (perception done)

> Perceived Regions
  RIGHT_WOMAN_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [REL_SVO_WHO_40]
[ @] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ @] SemRep-N KICK_20
[ @] SemRep-R PATIENT_21 from KICK_20 to WOMAN_23
[ @] SemRep-R AGENT_22 from KICK_20 to WOMAN_0
[!@] SemRep-N WOMAN_23
[ @] Construction SVO_24 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [WOMAN_11] [KICK_29] [REL_SVO_WHO_41]
[ O] Construction EXIST_S_26 covering WOMAN_23 for 'there is' [REL_SVO_WHO_41]
[ @] Construction KICK_29 covering KICK_20 for 'kick'
[!@] SemRep-N WEAR_30
[!@] SemRep-N DRESS_31
[!@] SemRep-R AGENT_32 from WEAR_30 to WOMAN_23
[!@] SemRep-R PATIENT_33 from WEAR_30 to DRESS_31
[!X] Construction SVO_35 covering WOMAN_23 DRESS_31 WEAR_30 AGENT_32 PATIENT_33 for [REL_PAS_SVO_WHO_42] [WEAR_43] [DRESS_45]
[!X] Construction PAS_SVO_36 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [REL_SVO_WHO_41] 'is' [KICK_29] '-ed by' [WOMAN_11]
[!X] Construction PAS_SVO_37 covering WOMAN_23 DRESS_31 WEAR_30 AGENT_32 PATIENT_33 for [DRESS_45] 'is' [WEAR_43] '-ed by' [REL_PAS_SVO_WHO_42]
[!O] Construction EXIST_S_39 covering DRESS_31 for 'there is' [DRESS_45]
[!X] Construction REL_SVO_WHO_40 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [WOMAN_11] 'who' [KICK_29] [REL_SVO_WHO_41]
[!@] Construction REL_SVO_WHO_41 covering WOMAN_23 DRESS_31 WEAR_30 AGENT_32 PATIENT_33 for [WOMAN_44] 'who' [WEAR_43] [DRESS_45]
[!X] Construction REL_PAS_SVO_WHO_42 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [WOMAN_44] 'who is' [KICK_29] '-ed by' [WOMAN_11]
[!@] Construction WEAR_43 covering WEAR_30 for 'wear'
[!@] Construction WOMAN_44 covering WOMAN_23 for 'woman'
[!@] Construction DRESS_45 covering DRESS_31 for 'dress'

> Competition Traces
  SVO_24(1424) eliminated PAS_SVO_36(668)
  SVO_24(1424) eliminated REL_SVO_WHO_40(714)
  SVO_24(1424) eliminated REL_PAS_SVO_WHO_42(718)
  SVO_35(718) eliminated PAS_SVO_37(662)
  REL_SVO_WHO_41(1424) eliminated SVO_35(718)

> Construction Structures
[ ] 138: EXIST_S_26 'there is' [WOMAN_44 'woman']
[ ] 138: EXIST_S_39 'there is' [DRESS_45 'dress']
[ ] 1036: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [WOMAN_44 'woman']
[X] 280: PAS_SVO_36 [WOMAN_44 'woman'] 'is' [KICK_29 'kick'] '-ed by' [WOMAN_11 'woman']
[X] 536: SVO_35 [WOMAN_44 'woman'] [WEAR_43 'wear'] [DRESS_45 'dress']
[X] 480: PAS_SVO_37 [DRESS_45 'dress'] 'is' [WEAR_43 'wear'] '-ed by' [WOMAN_44 'woman']
[X] 326: EXIST_S_7 'there is' [REL_SVO_WHO_40 [WOMAN_11 'woman'] 'who' [KICK_29 'kick'] [WOMAN_44 'woman']]
[X] 320: EXIST_S_26 'there is' [REL_PAS_SVO_WHO_42 [WOMAN_44 'woman'] 'who is' [KICK_29 'kick'] '-ed by' [WOMAN_11 'woman']]
[ ] 526: EXIST_S_26 'there is' [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [DRESS_45 'dress']]
[*] 1424: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [DRESS_45 'dress']]
[X] 668: PAS_SVO_36 [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [DRESS_45 'dress']] 'is' [KICK_29 'kick'] '-ed by' [WOMAN_11 'woman']
[X] 718: SVO_35 [REL_PAS_SVO_WHO_42 [WOMAN_44 'woman'] 'who is' [KICK_29 'kick'] '-ed by' [WOMAN_11 'woman']] [WEAR_43 'wear'] [DRESS_45 'dress']
[X] 662: PAS_SVO_37 [DRESS_45 'dress'] 'is' [WEAR_43 'wear'] '-ed by' [REL_PAS_SVO_WHO_42 [WOMAN_44 'woman'] 'who is' [KICK_29 'kick'] '-ed by'
[WOMAN_11 'woman']]
[X] 714: EXIST_S_7 'there is' [REL_SVO_WHO_40 [WOMAN_11 'woman'] 'who' [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear']
[DRESS_45 'dress']]]

> Produced Utterance
  "woman who wear dress"

> Next Attention
  RIGHT_DRESS_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 6
================================================================================
> Current Attention
  RIGHT_DRESS_AREA (perception done)

> Perceived Regions
  RIGHT_DRESS_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [ ]
[ @] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ @] SemRep-N KICK_20
[ @] SemRep-R PATIENT_21 from KICK_20 to WOMAN_23
[ @] SemRep-R AGENT_22 from KICK_20 to WOMAN_0
[ @] SemRep-N WOMAN_23
[ @] Construction SVO_24 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [WOMAN_11] [KICK_29] [REL_SVO_WHO_41]
[ O] Construction EXIST_S_26 covering WOMAN_23 for 'there is' [IN_COLOR_51]
[ @] Construction KICK_29 covering KICK_20 for 'kick'
[ @] SemRep-N WEAR_30
[ @] SemRep-N DRESS_31
[ @] SemRep-R AGENT_32 from WEAR_30 to WOMAN_23
[ @] SemRep-R PATIENT_33 from WEAR_30 to DRESS_31
[ O] Construction EXIST_S_39 covering DRESS_31 for 'there is' [ADJ_NOUN_50]
[ @] Construction REL_SVO_WHO_41 covering WOMAN_23 DRESS_31 WEAR_30 AGENT_32 PATIENT_33 for [WOMAN_44] 'who' [WEAR_43] [REL_SPA_WHICH_49]
```

```
[ @] Construction WEAR_43 covering WEAR_30 for 'wear'
[ @] Construction WOMAN_44 covering WOMAN_23 for 'woman'
[ @] Construction DRESS_45 covering DRESS_31 for 'dress'
[!@] SemRep-N BLUE_46
[!@] SemRep-R MODIFY_47 from BLUE_46 to DRESS_31
[!X] Construction SPA_48 covering DRESS_31 BLUE_46 MODIFY_47 for [DRESS_45] 'is' [BLUE_52]
[!@] Construction REL_SPA_WHICH_49 covering DRESS_31 BLUE_46 MODIFY_47 for [DRESS_45] 'which is' [BLUE_52]
[!X] Construction ADJ_NOUN_50 covering DRESS_31 BLUE_46 MODIFY_47 for [BLUE_52] [DRESS_45]
[!X] Construction IN_COLOR_51 covering WOMAN_23 WEAR_30 DRESS_31 BLUE_46 AGENT_32 PATIENT_33 MODIFY_47 for [WOMAN_44] 'in' [BLUE_52]
[!@] Construction BLUE_52 covering BLUE_46 for 'blue'

> Competition Traces
  REL_SVO_WHO_41(2313) eliminated IN_COLOR_51(632)
  REL_SPA_WHICH_49(2313) eliminated SPA_48(239)
  SPA_48(239) eliminated ADJ_NOUN_50(234)

> Construction Structures
[X] 239: SPA_48 [DRESS_45 'dress'] 'is' [BLUE_52 'blue']
[X] 632: EXIST_S_26 'there is' [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]
[ ] 227: EXIST_S_39 'there is' [REL_SPA_WHICH_49 [DRESS_45 'dress'] 'which is' [BLUE_52 'blue']]
[X] 234: EXIST_S_39 'there is' [ADJ_NOUN_50 [BLUE_52 'blue'] [DRESS_45 'dress']]
[X] 530: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [IN_COLOR_51 [WOMAN_44 'woman'] 'in' [BLUE_52 'blue']]
[ ] 115: EXIST_S_26 'there is' [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [REL_SPA_WHICH_49 [DRESS_45 'dress'] 'which is' [BLUE_52
'blue']]]
[X] 122: EXIST_S_26 'there is' [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [ADJ_NOUN_50 [BLUE_52 'blue'] [DRESS_45 'dress']]]
[*] 2313: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [REL_SPA_WHICH_49 [DRESS_45 'dress']
'which is' [BLUE_52 'blue']]]
[X] -80: SVO_24 [WOMAN_11 'woman'] [KICK_29 'kick'] [REL_SVO_WHO_41 [WOMAN_44 'woman'] 'who' [WEAR_43 'wear'] [ADJ_NOUN_50 [BLUE_52 'blue'] [DRESS_45
'dress']]]

> Produced Utterance
  "which is blue"

> Next Attention
  BOXINGRING_AREA (uncertainty left: 2)

================================================================================
  Simulation Time: 7
================================================================================
> Current Attention
  BOXINGRING_AREA (uncertainty left: 1)

> Schema Instances
[ x] SemRep-N WOMAN_0
[ O] Construction EXIST_S_7 covering WOMAN_0 for 'there is' [ ]
[ x] Construction WOMAN_11 covering WOMAN_0 for 'woman'
[ x] SemRep-N KICK_20
[ x] SemRep-R PATIENT_21 from KICK_20 to WOMAN_23
[ x] SemRep-R AGENT_22 from KICK_20 to WOMAN_0
[ x] SemRep-N WOMAN_23
[ x] Construction SVO_24 covering WOMAN_0 WOMAN_23 KICK_20 AGENT_22 PATIENT_21 for [ ] [ ] [ ]
[ O] Construction EXIST_S_26 covering WOMAN_23 for 'there is' [ ]
[ x] Construction KICK_29 covering KICK_20 for 'kick'
[ x] SemRep-N WEAR_30
[ x] SemRep-N DRESS_31
[ x] SemRep-R AGENT_32 from WEAR_30 to WOMAN_23
[ x] SemRep-R PATIENT_33 from WEAR_30 to DRESS_31
[ O] Construction EXIST_S_39 covering DRESS_31 for 'there is' [ ]
[ x] Construction REL_SVO_WHO_41 covering WOMAN_23 DRESS_31 WEAR_30 AGENT_32 PATIENT_33 for [ ] 'who' [ ] [ ]
[ x] Construction WEAR_43 covering WEAR_30 for 'wear'
[ x] Construction WOMAN_44 covering WOMAN_23 for 'woman'
[ x] Construction DRESS_45 covering DRESS_31 for 'dress'
[ x] SemRep-N BLUE_46
[ x] SemRep-R MODIFY_47 from BLUE_46 to DRESS_31
[ x] Construction REL_SPA_WHICH_49 covering DRESS_31 BLUE_46 MODIFY_47 for [ ] 'which is' [ ]
[ x] Construction BLUE_52 covering BLUE_46 for 'blue'

> Next Attention
  BOXINGRING_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 8
================================================================================
> Current Attention
  BOXINGRING_AREA (perception done)

> Perceived Regions
  BOXINGRING_AREA

> Schema Instances
[!@] SemRep-N BOXINGRING_53
[!X] SemRep-R IN_54 from ?? to BOXINGRING_53
[!@] Construction THEME_S_55 covering BOXINGRING_53 for 'it is' [BOXINGRING_56]
[!@] Construction BOXINGRING_56 covering BOXINGRING_53 for 'boxing ring'

> Construction Structures
[*] 136: THEME_S_55 'it is' [BOXINGRING_56 'boxing ring']

> Produced Utterance
  "it is boxing ring"

> Next Attention
  PEOPLE_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 9
================================================================================
> Current Attention
  PEOPLE_AREA (perception done)

> Perceived Regions
  PEOPLE_AREA

> Schema Instances
[ x] SemRep-N BOXINGRING_53
[ x] Construction THEME_S_55 covering BOXINGRING_53 for 'it is' [ ]
[ x] Construction BOXINGRING_56 covering BOXINGRING_53 for 'boxing ring'
```

```
[!@] SemRep-N PEOPLE_57
[!@] Construction EXIST_S_58 covering PEOPLE_57 for 'there is' [PEOPLE_59]
[!@] Construction PEOPLE_59 covering PEOPLE_57 for 'people'

> Construction Structures
[*] 137: EXIST_S_58 'there is' [PEOPLE_59 'people']

> Produced Utterance
  "there is people"

> Next Attention
  PEOPLE_FOCUS_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 10
================================================================================
> Current Attention
  PEOPLE_FOCUS_AREA (perception done)

> Perceived Regions
  PEOPLE_FOCUS_AREA

> Schema Instances
[ @] SemRep-N PEOPLE_57
[ @] Construction EXIST_S_58 covering PEOPLE_57 for 'there is' [REL_SV_WHO_64]
[ @] Construction PEOPLE_59 covering PEOPLE_57 for 'people'
[!@] SemRep-N WATCH_60
[!@] SemRep-R AGENT_61 from WATCH_60 to PEOPLE_57
[!X] SemRep-R CONCURRENT_62 from ?? to WATCH_60
[!X] Construction SV_63 covering PEOPLE_57 WATCH_60 AGENT_61 for [PEOPLE_59] [WATCH_65]
[!@] Construction REL_SV_WHO_64 covering PEOPLE_57 WATCH_60 AGENT_61 for [PEOPLE_59] 'who' [WATCH_65]
[!@] Construction WATCH_65 covering WATCH_60 for 'watch'

> Competition Traces
  REL_SV_WHO_64(529) eliminated SV_63(239)

> Construction Structures
[X] 239: SV_63 [PEOPLE_59 'people'] [WATCH_65 'watch']
[*] 529: EXIST_S_58 'there is' [REL_SV_WHO_64 [PEOPLE_59 'people'] 'who' [WATCH_65 'watch']]

> Produced Utterance
  "who watch"

> Next Attention
  None

================================================================================
  Simulation Time: 11
================================================================================
> Current Attention
  None

> Schema Instances
[ x] SemRep-N PEOPLE_57
[ x] Construction EXIST_S_58 covering PEOPLE_57 for 'there is' [ ]
[ x] Construction PEOPLE_59 covering PEOPLE_57 for 'people'
[ x] SemRep-N WATCH_60
[ x] SemRep-R AGENT_61 from WATCH_60 to PEOPLE_57
[ x] Construction REL_SV_WHO_64 covering PEOPLE_57 WATCH_60 AGENT_61 for [ ] 'who' [ ]
[ x] Construction WATCH_65 covering WATCH_60 for 'watch'

> Next Attention
  None

================================================================================
  Simulation Time: 12
================================================================================
> Current Attention
  None

> Next Attention
  None


Simulation complete: inactivity termination.
```

Figure 5.5-2 illustrated two cases of scene perception and description production. This appendix provides the simulation results corresponding to the structural case (A) and the incremental case (B). Threshold is set to “low” (the time parameter is tuned to “1”) for both of the cases.

The following is the scene description file used for the structural strategy (case A), in which the gist of the event (MOUSE-SQUIRT-TURTLE) is provided first.

```
#
# TCG Scene: structural strategy
#
# Kuchinsky's example of easy event with difficult objects
#


image: none
resolution: 0 * 0


region GIST
```

```
{
            location: 0, 0 size: 0, 0
            saliency: 0
            uncertainty: 0                      # instantly perceived

            perceive MOUSE=OBJECT, TURTLE=OBJECT, SQUIRT=ACTION
            perceive SQUIRT_AGENT, SQUIRT_PATIENT
}

region TURTLE_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 100                       # cued object
            uncertainty: 1

            object TURTLE { concept: TURTLE }

            perceive TURTLE
}

region SQUIRT_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 70
            uncertainty: 1

            object SQUIRT { concept: SQUIRT }
            relation SQUIRT_AGENT { concept: AGENT from: SQUIRT to: MOUSE }
            relation SQUIRT_PATIENT { concept: PATIENT from: SQUIRT to: TURTLE }

            perceive SQUIRT, SQUIRT_AGENT, SQUIRT_PATIENT
}

region MOUSE_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 50
            uncertainty: 1

            object MOUSE { concept: MOUSE }

            perceive MOUSE
}
```

The following is the simulation output for the structural strategy.

```
Template Construction Grammar (TCG) Simulator v2.5

Jinyong Lee (jinyongl@usc.edu), June 23. 2012
USC Brain Project, Computer Science Department
University of Southern California (USC)

Loading Initialization File 'tcg.ini'...
Loading Semantic Network 'TCG_semantics.txt'...
Loading Construction Vocabulary 'TCG_vocabulary_exp.txt'...
Loading Scene 'scene_structural.txt'...

Initializing Simulator...
- Max Simulation Time: 20
- Premature Production: on
- Utterance Continuity: on
- Verbal Guidance: on
- Threshold of Utterance: Time = 1, CNXs = infinite, Syllables = infinite

Beginning Simulation...

================================================================================
  Simulation Time: 1
================================================================================
> Current Attention
  None

> Perceived Regions
  GIST

> Schema Instances
[!O] SemRep-N OBJECT_0
[!O] SemRep-N OBJECT_1
[!O] SemRep-N ACTION_2
[!@] SemRep-R AGENT_3 from ACTION_2 to OBJECT_0
[!@] SemRep-R PATIENT_4 from ACTION_2 to OBJECT_1
[!@] Construction SVO_5 covering OBJECT_0 OBJECT_1 ACTION_2 AGENT_3 PATIENT_4 for [ ] [ ] [ ]
[!X] Construction PAS_SVO_6 covering OBJECT_0 OBJECT_1 ACTION_2 AGENT_3 PATIENT_4 for [ ] 'is' [ ] '-ed by' [ ]

> Competition Traces
  SVO_5(250) eliminated PAS_SVO_6(194)

> Construction Structures
[*] 250: SVO_5 [ ] [ ] [ ]
[X] 194: PAS_SVO_6 [ ] 'is' [ ] '-ed by' [ ]

> Produced Utterance
  "uh..."

> Next Attention
  MOUSE_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 2
================================================================================
> Current Attention
  MOUSE_AREA (perception done)

> Perceived Regions
  MOUSE_AREA
```

```
> Schema Instances
[!@] SemRep-N MOUSE_0
[ O] SemRep-N OBJECT_1
[ O] SemRep-N ACTION_2
[ @] SemRep-R AGENT_3 from ACTION_2 to MOUSE_0
[ @] SemRep-R PATIENT_4 from ACTION_2 to OBJECT_1
[ @] Construction SVO_5 covering MOUSE_0 OBJECT_1 ACTION_2 AGENT_3 PATIENT_4 for [MOUSE_9] [ ] [ ]
[!X] Construction PAS_SVO_8 covering MOUSE_0 OBJECT_1 ACTION_2 AGENT_3 PATIENT_4 for [ ] 'is' [ ] '-ed by' [MOUSE_9]
[!@] Construction MOUSE_9 covering MOUSE_0 for 'mouse'

> Competition Traces
  SVO_5(445) eliminated PAS_SVO_8(289)

> Construction Structures
[*] 445: SVO_5 [MOUSE_9 'mouse'] [ ] [ ]
[X] 289: PAS_SVO_8 [ ] 'is' [ ] '-ed by' [MOUSE_9 'mouse']

> Produced Utterance
  "mouse..."

> Next Attention
  SQUIRT_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 3
================================================================================
> Current Attention
  SQUIRT_AREA (perception done)

> Perceived Regions
  SQUIRT_AREA

> Schema Instances
[ @] SemRep-N MOUSE_0
[ O] SemRep-N OBJECT_1
[!@] SemRep-N SQUIRT_2
[!@] SemRep-R AGENT_3 from SQUIRT_2 to MOUSE_0
[!@] SemRep-R PATIENT_4 from SQUIRT_2 to OBJECT_1
[ @] Construction SVO_5 covering MOUSE_0 OBJECT_1 SQUIRT_2 AGENT_3 PATIENT_4 for [MOUSE_9] [SQUIRT_12] [ ]
[ @] Construction MOUSE_9 covering MOUSE_0 for 'mouse'
[!X] Construction PAS_SVO_11 covering MOUSE_0 OBJECT_1 SQUIRT_2 AGENT_3 PATIENT_4 for [ ] 'is' [SQUIRT_12] '-ed by' [MOUSE_9]
[!@] Construction SQUIRT_12 covering SQUIRT_2 for 'squirt at'

> Competition Traces
  SVO_5(737) eliminated PAS_SVO_11(281)

> Construction Structures
[*] 737: SVO_5 [MOUSE_9 'mouse'] [SQUIRT_12 'squirt at'] [ ]
[X] 281: PAS_SVO_11 [ ] 'is' [SQUIRT_12 'squirt at'] '-ed by' [MOUSE_9 'mouse']

> Produced Utterance
  "squirt at..."

> Next Attention
  TURTLE_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 4
================================================================================
> Current Attention
  TURTLE_AREA (perception done)

> Perceived Regions
  TURTLE_AREA

> Schema Instances
[ @] SemRep-N MOUSE_0
[!@] SemRep-N TURTLE_1
[ @] SemRep-N SQUIRT_2
[ @] SemRep-R AGENT_3 from SQUIRT_2 to MOUSE_0
[ @] SemRep-R PATIENT_4 from SQUIRT_2 to TURTLE_1
[ @] Construction SVO_5 covering MOUSE_0 TURTLE_1 SQUIRT_2 AGENT_3 PATIENT_4 for [MOUSE_9] [SQUIRT_12] [TURTLE_15]
[ @] Construction MOUSE_9 covering MOUSE_0 for 'mouse'
[ @] Construction SQUIRT_12 covering SQUIRT_2 for 'squirt at'
[!X] Construction PAS_SVO_14 covering MOUSE_0 TURTLE_1 SQUIRT_2 AGENT_3 PATIENT_4 for [TURTLE_15] 'is' [SQUIRT_12] '-ed by' [MOUSE_9]
[!@] Construction TURTLE_15 covering TURTLE_1 for 'turtle'

> Competition Traces
  SVO_5(1031) eliminated PAS_SVO_14(275)

> Construction Structures
[*] 1031: SVO_5 [MOUSE_9 'mouse'] [SQUIRT_12 'squirt at'] [TURTLE_15 'turtle']
[X] 275: PAS_SVO_14 [TURTLE_15 'turtle'] 'is' [SQUIRT_12 'squirt at'] '-ed by' [MOUSE_9 'mouse']

> Produced Utterance
  "turtle"

> Next Attention
  None

================================================================================
  Simulation Time: 5
================================================================================
> Current Attention
  None

> Schema Instances
[ x] SemRep-N MOUSE_0
[ x] SemRep-N TURTLE_1
[ x] SemRep-N SQUIRT_2
[ x] SemRep-R AGENT_3 from SQUIRT_2 to MOUSE_0
[ x] SemRep-R PATIENT_4 from SQUIRT_2 to TURTLE_1
[ x] Construction SVO_5 covering MOUSE_0 TURTLE_1 SQUIRT_2 AGENT_3 PATIENT_4 for [ ] [ ] [ ]
[ x] Construction MOUSE_9 covering MOUSE_0 for 'mouse'
[ x] Construction SQUIRT_12 covering SQUIRT_2 for 'squirt at'
[ x] Construction TURTLE_15 covering TURTLE_1 for 'turtle'
```

```
> Next Attention
  None

================================================================================
  Simulation Time: 6
================================================================================
> Current Attention
  None

> Next Attention
  None


Simulation complete: inactivity termination.
```

The following is the scene description file used for the incremental strategy (case B), in which the event is perceived in an incremental manner.

```
#
# TCG Scene: incremental strategy
#
# Kuchinsky's example of difficult event with easy objects
#


image: none
resolution: 0 * 0


region WOMAN_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 100                       # cued object
            uncertainty: 1

            object WOMAN { concept: WOMAN }

            perceive WOMAN
}

region TALK_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 70
            uncertainty: 1

            object TALK { concept: TALK }
            relation TALK_AGENT { concept: AGENT from: TALK to: WOMAN }
            relation TALK_PATIENT { concept: PATIENT from: TALK to: PEOPLE }

            perceive TALK, TALK_AGENT, TALK_PATIENT
            perceive PEOPLE=HUMAN
}

region PEOPLE_AREA
{
            location: 0, 0 size: 0, 0
            saliency: 50
            uncertainty: 1

            object PEOPLE { concept: PEOPLE }

            perceive PEOPLE
}
```

The following is the simulation output for the incremental strategy.

```
Template Construction Grammar (TCG) Simulator v2.5

Jinyong Lee (jinyongl@usc.edu), June 23. 2012
USC Brain Project, Computer Science Department
University of Southern California (USC)

Loading Initialization File 'tcg.ini'...
Loading Semantic Network 'TCG_semantics.txt'...
Loading Construction Vocabulary 'TCG_vocabulary_exp.txt'...
Loading Scene 'scene_incremental.txt'...

Initializing Simulator...
- Max Simulation Time: 20
- Premature Production: on
- Utterance Continuity: on
- Verbal Guidance: on
- Threshold of Utterance: Time = 1, CNXs = infinite, Syllables = infinite

Beginning Simulation...

================================================================================
  Simulation Time: 1
================================================================================
> Current Attention
  None

> Next Attention
  WOMAN_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 2
================================================================================
> Current Attention
```

```
  WOMAN_AREA (perception done)

> Perceived Regions
  WOMAN_AREA

> Schema Instances
[!@] SemRep-N WOMAN_0
[!@] Construction WOMAN_1 covering WOMAN_0 for 'woman'

> Construction Structures
[*] 95: WOMAN_1 'woman'

> Produced Utterance
  "woman"

> Next Attention
  TALK_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 3
================================================================================
> Current Attention
  TALK_AREA (perception done)

> Perceived Regions
  TALK_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ @] Construction WOMAN_1 covering WOMAN_0 for 'woman'
[!@] SemRep-N TALK_2
[!@] SemRep-R AGENT_3 from TALK_2 to WOMAN_0
[!@] SemRep-R PATIENT_4 from TALK_2 to HUMAN_5
[!O] SemRep-N HUMAN_5
[!@] Construction SVO_6 covering WOMAN_0 HUMAN_5 TALK_2 AGENT_3 PATIENT_4 for [WOMAN_1] [TALK_10] [ ]
[!X] Construction PAS_SVO_7 covering WOMAN_0 HUMAN_5 TALK_2 AGENT_3 PATIENT_4 for [ ] 'is' [TALK_10] '-ed by' [WOMAN_1]
[!X] Construction REL_SVO_WHO_8 covering WOMAN_0 HUMAN_5 TALK_2 AGENT_3 PATIENT_4 for [WOMAN_1] 'who' [TALK_10] [ ]
[!X] Construction REL_PAS_SVO_WHO_9 covering WOMAN_0 HUMAN_5 TALK_2 AGENT_3 PATIENT_4 for [ ] 'who is' [TALK_10] '-ed by' [WOMAN_1]
[!@] Construction TALK_10 covering TALK_2 for 'talk to'

> Competition Traces
  SVO_6(539) eliminated PAS_SVO_7(283)
  SVO_6(539) eliminated REL_SVO_WHO_8(486)
  SVO_6(539) eliminated REL_PAS_SVO_WHO_9(280)

> Construction Structures
[*] 539: SVO_6 [WOMAN_1 'woman'] [TALK_10 'talk to'] [ ]
[X] 283: PAS_SVO_7 [ ] 'is' [TALK_10 'talk to'] '-ed by' [WOMAN_1 'woman']
[X] 486: REL_SVO_WHO_8 [WOMAN_1 'woman'] 'who' [TALK_10 'talk to'] [ ]
[X] 280: REL_PAS_SVO_WHO_9 [ ] 'who is' [TALK_10 'talk to'] '-ed by' [WOMAN_1 'woman']

> Produced Utterance
  "talk to..."

> Next Attention
  PEOPLE_AREA (uncertainty left: 1)

================================================================================
  Simulation Time: 4
================================================================================
> Current Attention
  PEOPLE_AREA (perception done)

> Perceived Regions
  PEOPLE_AREA

> Schema Instances
[ @] SemRep-N WOMAN_0
[ @] Construction WOMAN_1 covering WOMAN_0 for 'woman'
[ @] SemRep-N TALK_2
[ @] SemRep-R AGENT_3 from TALK_2 to WOMAN_0
[ @] SemRep-R PATIENT_4 from TALK_2 to PEOPLE_5
[!@] SemRep-N PEOPLE_5
[ @] Construction SVO_6 covering WOMAN_0 PEOPLE_5 TALK_2 AGENT_3 PATIENT_4 for [WOMAN_1] [TALK_10] [PEOPLE_15]
[ @] Construction TALK_10 covering TALK_2 for 'talk to'
[!X] Construction PAS_SVO_12 covering WOMAN_0 PEOPLE_5 TALK_2 AGENT_3 PATIENT_4 for [PEOPLE_15] 'is' [TALK_10] '-ed by' [WOMAN_1]
[!X] Construction REL_SVO_WHO_13 covering WOMAN_0 PEOPLE_5 TALK_2 AGENT_3 PATIENT_4 for [WOMAN_1] 'who' [TALK_10] [PEOPLE_15]
[!X] Construction REL_PAS_SVO_WHO_14 covering WOMAN_0 PEOPLE_5 TALK_2 AGENT_3 PATIENT_4 for [PEOPLE_15] 'who is' [TALK_10] '-ed by' [WOMAN_1]
[!@] Construction PEOPLE_15 covering PEOPLE_5 for 'people'

> Competition Traces
  SVO_6(1033) eliminated PAS_SVO_12(277)
  SVO_6(1033) eliminated REL_SVO_WHO_13(280)
  SVO_6(1033) eliminated REL_PAS_SVO_WHO_14(274)

> Construction Structures
[*] 1033: SVO_6 [WOMAN_1 'woman'] [TALK_10 'talk to'] [PEOPLE_15 'people']
[X] 277: PAS_SVO_12 [PEOPLE_15 'people'] 'is' [TALK_10 'talk to'] '-ed by' [WOMAN_1 'woman']
[X] 280: REL_SVO_WHO_13 [WOMAN_1 'woman'] 'who' [TALK_10 'talk to'] [PEOPLE_15 'people']
[X] 274: REL_PAS_SVO_WHO_14 [PEOPLE_15 'people'] 'who is' [TALK_10 'talk to'] '-ed by' [WOMAN_1 'woman']

> Produced Utterance
  "people"

> Next Attention
  None

================================================================================
  Simulation Time: 5
================================================================================
> Current Attention
  None

> Schema Instances
[ x] SemRep-N WOMAN_0
[ x] Construction WOMAN_1 covering WOMAN_0 for 'woman'
[ x] SemRep-N TALK_2
```

```
[ x] SemRep-R AGENT_3 from TALK_2 to WOMAN_0
[ x] SemRep-R PATIENT_4 from TALK_2 to PEOPLE_5
[ x] SemRep-N PEOPLE_5
[ x] Construction SVO_6 covering WOMAN_0 PEOPLE_5 TALK_2 AGENT_3 PATIENT_4 for [ ] [ ] [ ]
[ x] Construction TALK_10 covering TALK_2 for 'talk to'
[ x] Construction PEOPLE_15 covering PEOPLE_5 for 'people'

> Next Attention
  None

================================================================================
  Simulation Time: 6
================================================================================
> Current Attention
  None

> Next Attention
  None


Simulation complete: inactivity termination.
```

References

Aboitiz, F., Aboitiz, S., & García, R. R. (2010). The Phonological Loop: A Key Innovation in Human Evolution. *Current Anthropology, 51*(1), S55-S65.

Aboitiz, F., García, R. R., Brunetti, E., & Bosman, C. (2006). The Origin of Broca's Area and its Connections from an Ancestral Working Memory Network. In Y. Grodzinsky & K. Amunts (Eds.), *Broca's region* (pp. 3-17). New York, NY: Oxford University Press.

Aboitiz, F., & García, R. V. (2009). Merging of phonological and gestural circuits in early language evolution. *Reviews in the Neurosciences, 20*(1), 71-84.

Aboitiz, F., & García V., R. (1997). The evolutionary origin of the language areas in the human brain. A neuroanatomical perspective. *Brain Research Reviews, 25*(3), 381-396.

Aggleton, J. P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences, 22*(3), 425-489.

Allen, K., Pereira, F., Botvinick, M., & Goldberg, A. E. (2012). Distinguishing Grammatical Constructions with fMRI Pattern Analysis. to appear.

Allum, P. H., & Wheeldon, L. R. (2007). Planning Scope in Spoken Sentence Production: The Role of Grammatical Units. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 33*(4), 791-810.

Almor, A., Smith, D. V., Bonilha, L., Fridriksson, J., & Rorden, C. (2007). What is in a name? Spatial brain circuits are used to track discourse references. *Neuroreport, 18*(12), 1215-1219.

Altmann, G. T. M. (2004). Language-mediated eye movements in the absence of a visual world: the 'blank screen paradigm'. *Cognition, 93*(2), B79-B87.

Amunts, K., Schleicher, A., Ditterich, A., & Zilles, K. (2003). Broca's region: Cytoarchitectonic asymmetry and developmental changes. *The Journal of Comparative Neurology, 465*(1), 72-89.

Andersson, B., Dahl, J., Holmqvist, K., Holsanova, J., Johansson, V., Karlsson, H., . . . Wengelin, A. (2006). Combining Keystroke Logging with Eye Tracking. In L. Waes, M. Leijten & C. M. Neuwirth (Eds.), *Writing and Digital Media* (pp. 166-172). North Holland: Elsevier.

Anwander, A., Tittgemeyer, M., von Cramon, D. Y., Friederici, A. D., & Knösche, T. R. (2007). Connectivity-Based Parcellation of Broca's area. *Cerebral Cortex, 17*(4), 816-825.

Arbib, M. A. (1972). *The Metaphorical Brain: An Introduction to Cybernetics as Artificial Intelligence and Brain Theory*. New York: Wiley Interscience.

Arbib, M. A. (1981). Perceptual structures and distributed motor control. In V. B. Brooks (Ed.), *Handbook of physiology - The nervous system. II. Motor control* (pp. 1449-1480): American Physiological Society.

Arbib, M. A. (1989). *The Metaphorical Brain 2: Neural Networks and Beyond*: Wiley-Interscience.

Arbib, M. A. (1995). Schema Theory. In M. A. Arbib (Ed.), The Handbook of Brain Theory and Neural Networks. Cambridge, MA: MIT Press.

Arbib, M. A. (2002). Schema Theory. In M. A. Arbib (Ed.), *The Handbook of Brain Theory and Neural Networks, 2nd Edition* (pp. 993-998). Cambridge, MA: MIT Press.

Arbib, M. A. (2006). The Mirror System Hypothesis on the linkage of action and languages. In M. A. Arbib (Ed.), *Action to Language: via the Mirror Neuron System* (pp. 3-47). Cambridge, UK: Cambridge University Press.

Arbib, M. A. (2008). From grasp to language: Embodied concepts and the challenge of abstraction. *Journal of Physiology-Paris, 102*(1-3), 4-20.

Arbib, M. A., & Caplan, D. (1979). Neurolinguistics must be computational. *Behavioral and Brain Sciences, 2*(3), 449-460.

Arbib, M. A., Conklin, E. J., & Hill, J. C. (1987). *From Schema Theory to Language*. New York: Oxford University Press.

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

Arbib, M. A., & Lee, J. (2007). Vision and Action in the Language-Ready Brain: From Mirror Neurons to SemRep. In F. Mele, G. Ramella, S. Santillo & F. Ventriglia (Eds.), *BVAI (Brain Vision & Artificial Intelligence) 2007, LNCS 4729* (pp. 104-123). Berlin/Heidelberg: Springer-Verlag.

Arbib, M. A., & Lee, J. (2008). Describing Visual Scenes: towards a neurolinguistics based on Construction Grammar. *Brain Research, 1225*, 146-162.

Arbib, M. A., & Liaw, J.-S. (1995). Sensorimotor transformations in the worlds of frogs and robots. *Artificial Intelligence, 72*(1-2), 53-79.

Awh, E., & Jonides, J. (2001). Overlapping mechanisms of attention and spatial working memory. *TRENDS in Cognitive Sciences, 5*(3), 119-126.

Awh, E., Vogel, E. K., & Oh, S.-H. (2006). Interactions between attention and working memory. *Neuroscience, 139*(1), 201-208.

Baddeley, A. D. (1996). *Working Memory*. New York: Oxford University Press.

Baddeley, A. D. (2003). Working memory: looking back and looking forward. *Nature Reviews. Neuroscience, 4*(10), 829-839.

Badre, D., & Wagner, A. D. (2007). Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia, 45*(13), 2883-2901.

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: A dynamic account of the N400. *Language and Cognitive Processes, 0*(0), 1-30.

Ballard, D. H., Hayhoe, M. M., & Pelz, J. B. (1995). Memory Representations in Natural Tasks. *Journal of Cognitive Neuroscience, 7*(1), 66-80.

Barrès, V., & Lee, J. (2013). Template Construction Grammar: From Visual Scene Description to Language Comprehension and Agrammatism. *Neuroinformatics, Published Online*.

Barsalou, L. W. (1999). Perceptual Symbol Systems. *Behavioral and Brain Sciences, 22*(4), 577-+.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *TRENDS in Cognitive Sciences, 7*(2), 84-91.

Beauchamp, M. S., & Martin, A. (2007). Grounding object concepts in perception and action: Evidence from fMRI studies of tools. *Cortex, 43*(3), 461-468.

Belopolsky, A. V., & Theeuwes, J. (2009). Inhibition of saccadic eye movements to locations in spatial working memory. *Attention, Perception, & Psychophysics, 71*(3), 620-631.

Bergen, B. K., & Chang, N. (2005). Embodied construction grammar in simulation-based language understanding. In J.-O. Ostman & M. Fried (Eds.), *Construction grammars: cognitive grounding and theoretical extensions* (pp. 147-190). Amsterdam: John Benjamins.

Bergen, B. K., & Wheeler, K. B. (2010). Grammatical aspect and mental simulation. *Brain & Language, 112*(3), 150-158.

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *TRENDS in Cognitive Sciences, 15*(11), 527-536.

Binkofski, F., Amunts, K., Stephan, K. M., Posse, S., Schormann, T., Freund, H.-J., . . . Seitz, R. J. (2000). Broca's region subserves imagery of motion: A combined cytoarchitectonic and fMRI study. *Human Brain Mapping, 11*(4), 273-285.

Binkofski, F., Buccino, G., Stephan, K. M., Rizzolatti, G., Seitz, R. J., & Freund, H.-J. (1999). A parieto-premotor network for object manipulation: evidence from neuroimaging. *Experimental Brain Research, 128*(1-2), 210-213.

Bisley, J. W., & Goldberg, M. E. (2003). Neuronal Activity in the Lateral Intraparietal Area and Spatial Attention. *Science, 299*(5603), 81-86.

Blumenfeld, R. S., & Ranganath, C. (2006). Dorsolateral Prefrontal Cortex Promotes Long-Term Memory Formation through Its Role in Working Memory Organization. *The Journal of Neuroscience, 26*(3), 916-925.

Bock, K., & Cutting, C. (1992). Regulating mental energy: Performance units in language production. *Journal of Memory and Language, 31*(1), 99-127.

Bock, K., Irwin, D. E., & Davidson, D. J. (2004). Putting First Things First. In J. M. Henderson & F. Ferreira (Eds.), *The Interface of Language, Vision, and Action* (pp. 249-278). New York, NY: Psychology Press.

Bock, K., Irwin, D. E., Davidson, D. J., & Levelt, W. J. M. (2003). Minding the clock. *Journal of Memory and Language, 48*(4), 653-685.

Bock, K., & Levelt, W. J. M. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), Handbook of Psycholinguistics. San Diego, CA: Academic Press.

Bornkessel-Schlesewsky, I., & Schlesewsky, M. (2008). An alternative perspective on "semantic P600" effects in language comprehension. *Brain Research Reviews, 59*(1), 55-73.

Bornkessel, I., & Schlesewsky, M. (2006). The Extended Argument Dependency Model: A Neurocognitive Approach to Sentence Comprehension Across Languages. *Psychological Review, 113*(4), 787-821.

Brown-Schmidt, S., & Tanenhaus, M. K. (2006). Watching the eyes when talking about size: An investigation of message formulation and utterance planning. *Journal of Memory and Language, 54*(4), 592-609.

Buccino, G., Binkofski, F., Fink, G. R., Fadiga, L., Fogassi, L., Gallese, V., . . . Freund, H.-J. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *European Journal of Neuroscience, 13*(2), 400-404.

Buchsbaum, B. R., Olsen, R. K., Koch, P., & Berman, K. F. (2005). Human Dorsal and Ventral Auditory Streams Subserve Rehearsal-Based and Echoic Processes during Verbal Working Memory. *Neuron, 48*(4), 687-697.

Buffalo, E. A., Reber, P. J., & Squire, L. R. (1998). The human perirhinal cortex and recognition memory. *Hippocampus, 8*(4),

330-339.

Burgess, N., Becker, S., King, J. A., & O'Keefe, J. (2001). Memory for events and their spatial context: models and experiments. *Philosophical Transactions of the Royal Society of Biological Sciences, 356*(1413), 1493-1503.

Caramazza, A., & Shelton, J. R. (1998). Domain-Specific Knowledge Systems in the Brain: The Animate-Inanimate Distinction. *Journal of Cognitive Neuroscience, 10*(1), 1-34.

Carrasco, M., & Yeshurun, Y. (2009). Covert attention effects on spatial resolution. *Progress in Brain Research, 176*(9), 65-86.

Castiello, U., & Umilta, C. (1990). Size of the attentionalfocus and efficiency of processing. *Acta Psychologica, 73*(3), 195-209.

Catani, M., Jones, D. K., & Ffytche, D. H. (2005). Perisylvian Language Networks of the Human Brain. *Annals of Neurology, 57*(1), 8-16.

Cavanagh, P., & Alvarez, G. A. (2005). Tracking multiple targets with multifocal attention. *TRENDS in Cognitive Sciences, 9*(7), 349-354.

Cave, K. R., & Bichot, N. P. (1999). Visuospatial attention: Beyond a spotlight model. *Psychonomic Bulletin & Review, 6*(2), 204-223.

Cave, K. R., & Kosslyn, S. M. (1989). Varieties of size-specific visual selection. *Journal of Experimental Psychology: General, 118*(2), 148-164.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming Syntactic. *Psychological Review, 113*(2), 234-272.

Chee, M. W. L., Sriram, N., Soon, C. S., & Lee, K. M. (2000). Dorsolateral prefrontal cortex and the implicit association of concepts and attributes. *Neuroreport, 11*(1), 135-140.

Chein, J. M., & Fiez, J. A. (2001). Dissociation of Verbal Working Memory System Components Using a Delayed Serial Recall Task. *Cerebral Cortex, 11*(11), 1003-1014.

Chen, Z. (2003). Attentional focus, processing load, and Stroop interference. *Perception & Psychophysics, 65*(6), 888-900.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1970). Remarks on nominalization. In R. A. Jacobs & P. S. Rosenbaum (Eds.), *Readings in English transformational grammar* (pp. 184-221). Waltham: Ginn.

Chomsky, N. (1991). Some notes on economy of derivation and representation. In R. Freidin (Ed.), *Principles and Parameters in Comparative Grammar* (pp. 417-454). Cambridge, MA: MIT Press.

Chomsky, N. (1995). *The minimalist program*. Cambridge, MA: MIT Press.

Chouinard, P. A., & Goodale, M. A. (2010). Category-specific neural processing for naming pictures of animals and naming pictures of tools: An ALE meta-analysis. *Neuropsychologia, 48*(2), 409-418.

Christianson, K., & Ferreira, F. (2005). Conceptual accessibility and sentence production in a free word order language (Odawa). *Cognition, 98*(2), 105-135.

Chun, M. M., & Potter, M. C. (1995). A Two-Stage Model for Multiple Target Detection in Rapid Serial Visual Presentation. *Journal of Experimental Psychology: Human Perception and Performance, 21*(1), 109-127.

Colby, C. L., Duhamel, J.-R., & Goldberg, M. E. (1995). Oculocentric Spatial Representation in Parietal Cortex. *Cerebral*

Cortex, 5(5), 470-481.

Colby, C. L., & Goldberg, M. E. (1999). Space and attention in parietal cortex. *Annual Review of Neuroscience, 22*(1), 319-349.

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., & Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience, 3*(3), 292-297.

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*(5), 570-581.

Coslett, H. B., & Saffran, E. (1991). Simultanagnosia: To see but not two see. *Brain, 114*(4), 1523-1545.

Cowan, N. (1997). *Attention and Memory: An Integrated Framework* (Vol. 1). New York, NY: Oxford University Press.

Cowan, N. (1999). An Embedded-Process Model of Working Memory. In A. Miyake & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 62-101). Cambridge: Cambridge University Press.

Cowan, N. (2000). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87-185.

Croft, W. A. (2001). *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford: Oxford University Press.

Croft, W. A., & Cruse, D. A. (2005). *Cognitive Linguistics*. Cambridge: Cambridge University Press.

Curtis, C. E. (2006). Prefrontal and parietal contributions to spatial working memory. *Neuroscience, 139*(1), 173-180.

Curtis, C. E., & D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *TRENDS in Cognitive Sciences, 7*(9), 415-423.

Curtis, C. E., Sun, F. T., Miller, L. M., & D'Esposito, M. (2005). Coherence between fMRI time-series distinguishes two spatial working memory networks. *NeuroImage, 26*(1), 177-183.

D'Esposito, M., Postle, B. R., Ballard, D. H., & Lease, J. (1999). Maintenance versus Manipulation of Information Held in Working Memory: An Event-Related fMRI Study. *Brain and Cognition, 41*(1), 66-86.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition, 33*(1), 25-62.

Damasio, A. R., & Tranel, D. (1993). Nouns and verbs are retrieved with differently distributed neural systems. *Proceedings of the National Academy of Sciences of the United States of America, 90*(11), 4957-4960.

Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., & Damasio, A. R. (1996). A neural basis for lexical retrieval. *Nature, 380*(6574), 499-505.

Damasio, H., Tranel, D., Grabowski, T. J., Adolphs, R., & Damasio, A. R. (2004). Neural systems behind word and concept retrieval. *Cognition, 92*(1-2), 179-229.

De Beule, J., & Steels, L. (2005). *Hierarchy in Fluid Construction Grammar.* Paper presented at the Advances in Artificial Intelligence, Koblenz, Germany.

de Fockert, J. W., Rees, G., Frith, C. D., & Lavie, N. (2001). The Role of Working Memory in Visual Selective Attention. *Science, 291*(5509), 1803-1806.

Deco, G., & Schürmann, B. (2000). A hierarchical neural system with attentional top-down enhancement of the spatial resolution for object recognition. *Vision Research, 40*(20), 2845-2859.

Dehaene, S., & Cohen, L. (1994). Dissociable mechanisms of subitizing and counting: neuropsychological evidence from simultanagnosic patients. *Journal of Experimental Psychology: Human Perception and Performance, 20*(5), 958-975.

Desai, R. H., Binder, J. R., Conant, L. L., Mano, Q. R., & Seidenberg, M. S. (2011). The Neural Career of Sensory-motor Metaphors. *Journal of Cognitive Neuroscience, 23*(9), 2376-2386.

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of Biological Sciences, 353*(1373), 1245-1255.

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience, 18*(1), 193-222.

Deubel, H., Irwin, D. E., & Schneider, W. X. (1999). The subjective direction of gaze shifts long before the saccade. In W. Becker, H. Deubel & T. Mergner (Eds.), *Current oculomotor research: physiological and psychological aspects* (pp. 65-70). New York, NY: Plenum Publishers.

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*(12), 1827-1837.

Didday, R. L., & Arbib, M. A. (1975). Eyemovements and visualperception: A "two visual system" model. *International Journal of Man-Machine Studies, 7*(4), 547-569.

Diessel, H., & Tomasello, M. (2001). The acquisition of finite complement clauses in English: A corpus-based analysis. *Cognitive Linguistics, 12*(2), 97-142.

Dominey, P. F., & Arbib, M. A. (1992). A Cortico-Subcortical Model for Generation of Spatially Accurate Sequential Saccades. *Cerebral Cortex, 2*(2), 153-175.

Dominey, P. F., Hoen, M., & Inui, T. (2006). A Neurolinguistic Model of Grammatical Construction Processing. *Journal of Cognitive Neuroscience, 18*(12), 2088-2107.

Dominey, P. F., Inui, T., & Hoen, M. (2009). Neural network processing of natural language: II. Towards a unified model of corticostriatal function in learning sentence comprehension and non-linguistic sequencing. *Brain & Language, 109*(2), 80-92.

Draper, B. A., Collins, R. T., Brolio, J., Hanson, A. R., & Riseman, E. M. (1989). The schema system. *International Journal of Computer Vision, 2*(3), 209-250.

Dronkers, N. F., Wilkins, D. P., Jr., R. D. V. V., Redfern, B. B., & Jaeger, J. J. (2004). Lesion analysis of the brain areas involved in language comprehension. *Cognition, 92*(1-2), 145-177.

Duncan, J. (1984). Selective Attention and the Organization of Visual Information. *Journal of Experimental Psychology: General, 113*(4), 501-517.

Duncan, J. (2006). EPS Mid-career Award 2004: Brain Mechanisms of attention. *The Quarterly Journal of Experimental Psychology, 59*(1), 2-27.

Emmorey, K. (2002). *Language, cognition, and the brain: insights from sign language research*. Mahwah, NJ: Lawrence

Erlbaum Associates.

Epstein, R., Graham, K. S., & Downing, P. E. (2003). Viewpoint-Specific Scene Representations in Human Parahippocampal Cortex. *Neuron, 37*(5), 865-876.

Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The Parahippocampal Place Area: Recognition, Navigation, or Encoding? *Neuron, 23*(1), 115-125.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature, 392*(6676), 598-601.

Ericsson, K. A., & Kintsch, W. (1995). Long-Term Working Memory. *Psychological Review, 102*(2), 211-245.

Eriksen, C. W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics, 40*(4), 225-240.

Eriksen, C. W., & Yeh, Y.-Y. (1985). Allocation of Attention in the Visual Field. *Journal of Experimental Psychology: Human Perception and Performance, 11*(5), 583-597.

Erman, L. D., Hayes-Roth, F., Lesser, V. R., & Reddy, D. R. (1980). The HEARSAY-II Speech-Understanding System: Integrating Knowledge to Resolve Uncertainty. *ACM Computing Surveys, 12*(2), 213-253.

Evans, V., & Green, M. (2006). *Cognitive Linguistics: An Introduction*. New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

Fagg, A. H., & Arbib, M. A. (1998). Modeling parietal-premotor interactions in primate control of grasping. *Neural Networks, 11*(7-8), 1277-1303.

Farah, M. J. (1990). *Visual agnosia. Disorders of object recognition and what they tell us about normal vision*. Cambridge, MA: MIT Press.

Fei-Fei, L., Iyer, A., Koch, C., & Perona, P. (2007). What do we perceive in a glance of a real-world scene? *Journal of Vision, 7*(1), article 10.

Ferreira, F. (1993). Creation of prosody during sentence production. *Psychological Review, 100*(2), 233-253.

Ferreira, F., & Swets, B. (2002). How Incremental Is Language Production? Evidence from the Production of Utterances Requiring the Computation of Arithmetic Sums. *Journal of Memory and Language, 46*(1), 57-84.

Fiebach, C. J., Schlesewsky, M., & Friederici, A. D. (2001). Syntactic Working Memory and the Establishment of Filler-Gap Dependencies: Insights from ERPs and fMRI. *Journal of Psycholinguistic Research, 30*(3), 321-338.

Fiebach, C. J., Schlesewsky, M., Lohmann, G., von Cramon, D. Y., & Friederici, A. D. (2005). Revisiting the Role of Broca's Area in Sentence Processing: Syntactic Integration versus Syntactic Working Memory. *Human Brain Mapping, 24*(2), 79-91.

Fillmore, C. J., Kay, P., & O'Connor, M. C. (1988). Regularity and Idiomaticity in Grammatical Constructions: The case fo Let Alone. *Language, 64*(3), 501-538.

Fischer, M. H., & Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *The Quarterly Journal of Experimental Psychology, 61*(6), 1-26.

Flores d'Arcais, G. B. (1975). Some perceptual determinants of sentence construction. In G. B. Flores d'Arcais (Ed.), *Studies in perception: Festschrift for Fabio Metelli* (pp. 344-373). Milan, Italy: Martello-Giunti.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong*. New York, NY: Clarendon Press.

Fougnie, D., & Marois, R. (2007). Executive working memory load induces inattentional blindness. *Psychonomic Bulletin & Review, 14*(1), 142-147.

Frey, S., Campbell, J. S. W., Pike, G. B., & Petrides, M. (2008). Dissociating the Human Language Pathways with High Angular Resolution Diffusion Fiber Tractography. *The Journal of Neuroscience, 28*(45), 11435-11444.

Friederici, A. D. (2002). Towards a neural basis of auditory sentence processing. *TRENDS in Cognitive Sciences, 6*(2), 78-84.

Friederici, A. D. (2006). Broca's Area and the Ventral Premotor Cortex in Language: Functional Differentiation and Specificity. *Cortex, 42*(4), 472-475.

Friederici, A. D. (2009). Pathways to language: fiber tracts in the human brain. *TRENDS in Cognitive Sciences, 13*(4), 175-181.

Friederici, A. D., Opitz, B., & von Cramon, D. Y. (2000). Segregating Semantic and Syntactic Aspects of Processing in the Human Brain: an fMRI Investigation of Different Word Types. *Cerebral Cortex, 10*(7), 698-705.

Friederici, A. D., Rüschemeyer, S.-A., Hahne, A., & Fiebach, C. J. (2003). The Role of Left Inferior Frontal and Superior Temporal Cortex in Sentence Comprehension: Localizing Syntactic and Semantic Processes. *Cerebral Cortex, 13*(2), 170-177.

Fuster, J. M. (1997). Network memory. *Trends in Neuroscience, 20*(10), 451-459.

Fuster, J. M., Bodner, M., & Kroger, J. K. (2000). Cross-modal and cross-temporal association in neurons of frontal cortex. *Nature, 405*(6784), 347-351.

Gainotti, G. (2000). What the Locus of Brain Lesion Tells us About the Nature of the Cognitive Defect Underlying Category-Specific Disorders: A Review. *Cortex, 36*(4), 539-559.

Gallese, V., & Lakoff, G. (2005). The Brain's concepts: the role of the Sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology, 22*(3-4), 455-479.

Garavan, H. (1998). Serial attention within working memory. *Memory & Cognition, 26*(2), 263-276.

Gibson, J. J. (1986). *The Ecological Approach to Visual Perception*. New York: Psychology Press.

Gleitman, L. R., January, D., Nappa, R., & Trueswell, J. C. (2007). On the give and take between even apprehension and utterance formulation. *Journal of Memory and Language, 57*(4), 544-569.

Glenberg, A. M. (1997). What memory is for. *Behavioral and Brain Sciences, 20*(1), 1-55.

Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review, 9*(3), 558-565.

Gold, B. T., Balota, D. A., Jones, S. J., Powell, D. K., Smith, C. D., & Andersen, A. H. (2006). Dissociation of Automatic and Strategic Lexical-Semantics: Functional Magnetic Resonance Imaging Evidence for Differing Roles of Multiple Frontotemporal Regions. *The Journal of Neuroscience, 26*(24), 6523-6532.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago, IL: Chicago University Press.

Goldberg, A. E. (2003). Constructions: a new theoretical approach to language. *TRENDS in Cognitive Sciences, 7*(5), 219-224.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford: Oxford University Press.

Goodale, M. A., Jakobson, L. S., & Keillor, J. M. (1994). Differences in the visual control of pantomimed and natural

grasping movements. *Neuropsychologia, 32*(10), 1159-1178.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neuroscience, 15*(1), 20-25.

Goodale, M. A., Milner, A. D., Jakobson, L. S., & Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature, 349*(6305), 154-156.

Gordon, P. (2003). *The Origin of Argument Structure in Infant Event Representations.* Paper presented at the Proceedings of the 28th annual Boston University Conference on Language Development, Boston, MA.

Gottlieb, J. P., Kusunoki, M., & Goldberg, M. E. (1998). The representation of visual salience in monkey parietal cortex. *Nature, 391*(6666), 481-484.

Greene, M. R., & Oliva, A. (2009). The Briefest of Glances: The Time Course of Natural Scene Understanding. *Psychological Science, 20*(4), 464-472.

Greeno, J. G. (1994). Gibson's Affordances. *Psychological Review, 101*(2), 336-342.

Griffin, Z. M. (2001). Gaze durations during speech reflect word selection and phonological encoding. *Cognition, 82*(1), B1-B14.

Griffin, Z. M. (2003). A reversed word length effect in coordinating the preparation and articulation of words in speaking. *Psychonomic Bulletin & Review, 10*(3), 603-609.

Griffin, Z. M. (2004). Why Look? Reasons for Eye Movements Related to Language Production. In J. M. Henderson & F. Ferreira (Eds.), *The Interface of Language, Vision, and Action* (pp. 213-248). New York, NY: Psychology Press.

Griffin, Z. M., & Bock, K. (2000). What the Eyes Say About Speaking. *Psychological Science, 11*(4), 274-279.

Griffin, Z. M., & Garton, K. L. (2003). *Procrastination in Speaking: Ordering Arguments During Speech*. Paper presented at the The 16th Annual CUNY Conference on Human Sentence Processing, Boston, MA. Poster retrieved from

Griffin, Z. M., & Mouzon, S. (2004). *Can speakers order a sentence's arguments while saying it?* Paper presented at the The 17th Annual CUNY Sentence Processing Conference, College Park, MD.

Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research, 41*(10-11), 1409-1422.

Grill-Spector, K., & Malach, R. (2004). The Human Visual Cortex. *Annual Review of Neuroscience, 27*, 649-677.

Grodzinsky, Y. (2000). The neurology of syntax: Language use without Broca's area. *Behavioral and Brain Sciences, 23*(1), 1-71.

Hagiwara, H. (1993). The Breakdown of Japanese Passives and Theta-Role Assignment Principle by Broca's Aphasics. *Brain & Language, 45*(3), 318-339.

Hagoort, P. (2005). On Broca, brain, and binding: a new framework. *TRENDS in Cognitive Sciences, 9*(9), 416-423.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of Word Meaning and World Knowledge in Language Comprehension. *Science, 304*(5669), 438-441.

Hanson, S. J., Hanson, C., Halchenko, Y., Matsuka, T., & Zaimi, A. (2007). Bottom-up and top-down brain functional connectivity underlying comprehension of everyday visual action. *Brain Structure and Function, 212*(3-4), 231-244.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic Representation of Action Words in Human Motor and

Premotor Cortex. *Neuron, 41*(2), 301-307.

Hayhoe, M. M., Bensinger, D. G., & Ballard, D. H. (1998). Task Constraints in Visual Working Memory. *Vision Research, 38*(1), 125-137.

Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Theory*. New York: Wiley.

Heim, S., Eickhoff, S. B., & Amunts, K. (2009). Different roles of cytoarchitectonic BA 44 and BA 45 in phonological and semantic verbal fluency as revealed by dynamic causal modelling. *NeuroImage, 48*(3), 616-624.

Henderson, J. M. (1994). Two representational systems in dynamic visual identification. *Journal of Experimental Psychology: General, 123*(4), 410-426.

Henderson, J. M., & Ferreira, F. (2004). *The Interface of Language, Vision, and Action*. New York, NY: Psychology Press.

Henderson, J. M., & Hollingworth, A. (1999). High-level Scene Perception. *Annual Review of Psychology, 50*(1), 243-271.

Henson, R., Hartley, T., Burgess, N., Hitch, G., & Flude, B. (2003). Selective interference with verbal short-term memory for serial order information: a new paradigm and tests of a timing-signal hypothesis. *Quarterly Journal of Experimental Psychology, 56*(8), 1307-1334.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience, 8*(5), 393-402.

Hill, J. C. (1983). A computational model of language acquisition in the two-year-old. *Cognition and Brain Theory, 6*, 287-317.

Hochstein, S., & Ahissar, M. (2002). View from the Top: Hierarchies and Reverse Hierarchies in the Visual System. *Neuron, 36*(5), 791-804.

Hoen, M., Pachot-Clouard, M., Segebarth, C., & Dominey, P. F. (2006). When Broca Experiences the Janus Syndrome: an ER-FMRI Study Comparing Sentence Comprehension and Cognitive Sequence Processing. *Cortex, 42*(4), 605-623.

Hollingworth, A., & Henderson, J. M. (1999). Objectidentification is isolated from scenesemantic constraint: evidence from objecttype and tokendiscrimination. *Acta Psychologica, 102*(2-3), 319-343.

Holsanova, J. (2008). *Discourse, Vision, and Cognition*: John Benjamins Publishing Company.

Hopf, J.-M., Luck, S. J., Boelmans, K., Schoenfeld, M. A., Boehler, C. N., Rieger, J., & Heinze, H.-J. (2006). The Neural Site of Attention Matches the Spatial Scale of Perception. *The Journal of Neuroscience, 26*(13), 3532-3540.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America, 79*(8), 2554-2558.

Horowitz, T. S., Birnkrant, R. S., Fencsik, D. E., Tran, L., & Wolfe, J. M. (2006). How do we track invisible objects? *Psychonomic Bulletin & Review, 13*(3), 516-523.

Horowitz, T. S., Holcombe, A. O., Wolfe, J. M., Arsenio, H. C., & DiMase, J. S. (2004). Attentional pursuit is faster than attentional saccade. *Journal of Vision, 4*(7), 585-603.

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*(14), 1868-1876.

Ingle, D., Schneider, G., Trevarthen, C., & Held, R. (1967). Locating and identifying: Two modes of visual processing.

Psychological Research, 31(1), 42-43.

Iriki, A., & Taoka, M. (2012). Triadic (ecological, neural, cognitive) niche construction: a scenario of human brain evolution extrapolating tool use and language from the control of reaching actions. *Philosophical Transactions of the Royal Society of Biological Sciences, 367*(1585), 10-23.

Irwin, D. E. (2004). Fixation Location and Fixation Duration as Indices of Cognitive Processing. In J. M. Henderson & F. Ferreira (Eds.), *The Interface of Language, Vision, and Action* (pp. 105-133). New York, NY: Psychology Press.

Irwin, D. E., & Zelinsky, G. J. (2002). Eye movements and scene perception: Memory for things observed. *Attention, Perception, & Psychophysics, 64*(6), 882-895.

Ishai, A., Ungerleider, L. G., Martin, A., & Haxby, J. V. (2000). The Representation of Objects in the Human Occipital and Temporal Cortex. *Journal of Cognitive Neuroscience, 12*(2), 35-51.

Israel, M., Johnson, C., & Brooks, P. J. (2000). From states to events: The acquisition of English passive participles. *Cognitive Linguistics, 11*(1-2), 103-129.

Itti, L., & Arbib, M. A. (2006). Attention and the minimal subscene. In M. A. Arbib (Ed.), *Action to Language: via the Mirror Neuron System* (pp. 289-346). Cambridge, UK: Cambridge University Press.

Iwata, S. (2005). Locative alternation and two levels of verb meaning. *Cognitive Linguistics, 16*(2), 355-407.

Jackendoff, R. (1977). *X-bar-Syntax: A Study of Phrase Structure*. Cambridge, MA: MIT Press.

Jacobs, R. A., & Kosslyn, S. M. (1994). Encoding Shape and Spatial Relations: The Role of Receptive Field Size in Coordinating Complementary Representations *Cognitive Science, 18*(3), 361-386.

Jacquemot, C., & Scott, S. K. (2006). What is the relationship between phonological short-term memory and speech processing. *TRENDS in Cognitive Sciences, 10*(11), 480-486.

Jager, G., & Postma, A. (2003). On the hemispheric specialization for categorical and coordinate spatial relations: a review of the current evidence. *Neuropsychologia, 41*(4), 504-515.

Jeannerod, M., Decety, J., & Michel, F. (1994). Impairment of grasping movements following a bilateral posterior parietal lesion. *Neuropsychologia, 32*(4), 369-380.

Jefferies, L. N., Gmeindl, L., & Yantis, S. (2011). The distribution of visuospatial attention is influenced by illusory differences in the size of physically identical objects. *Journal of Vision, 11*(11), article 237.

Jones, D. M., Farrand, P., Stuart, G., & Morris, N. (1995). Functional equivalence of verbal and spatial information in serial short-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 1008-1018.

Kaan, E., & Swaab, T. Y. (2002). The brain circuitry of syntactic comprehension. *TRENDS in Cognitive Sciences, 6*(8), 350-356.

Kahneman, D., Treisman, A. M., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology, 24*(2), 175-219.

Kan, I. P., Barsalou, L. W., Solomon, K. O., Minor, J. K., & Thompson-Schill, S. L. (2003). Role of mental imagery in a property verification task: FMRI evidence for perceptual representations of conceptual knowledge. *Cognitive Neuropsychology, 20*(3), 525-540.

Kane, M. J., & Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and

general fluid intelligence: An individual-differences perspective. *Psychonomic Bulletin & Review, 9*(4), 637-671.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The Generality of Working Memory Capacity: A Latent-Variable Approach to Verbal and Visuospatial Memory Span and Reasoning. *Journal of Experimental Psychology: General, 133*(2), 189-217.

Karnath, H.-O. (2001). New insights into the functions of the superior temporal cortex. *Nature Reviews Neuroscience, 2*(8), 568-576.

Karnath, H.-O., Ferber, S., Rorden, C., & Driver, J. (2000). The fate of global information in dorsal simultanagnosia. *Neurocase, 6*(4), 295-306.

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron, 22*(4), 751-761.

Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of Visual Attention in the Human Cortex. *Annual Review of Neuroscience, 23*(1), 315-341.

Kawashima, R., Naitoh, E., Matsumura, M., Itoh, H., Ono, S., Satoh, K., . . . Fukuda, H. (1996). Topographic representation in human intraparietal sulcus of reaching and saccade. *Neuroreport, 7*(7), 1253-1256.

Keller, T. A., Carpenter, P. A., & Just, M. A. (2001). The Neural Bases of Sentence Comprehension: a fMRI Examination of Syntactic and Lexical Processing. *Cerebral Cortex, 11*(3), 223-237.

Kemmerer, D. (2000). Grammatically relevant and grammatically irrelevant features of verb meaning can be independently impaired. *Aphasiology, 14*(10), 997-1020.

Kemmerer, D. (2003). Why can you hit someone on the arm but not break someone on the arm? A neuropsychological investigation of the English body-part possessor ascension construction. *Journal of Neurolinguistics, 16*(1), 13-36.

Kemmerer, D. (2006). Action verbs and argument structure constructions. In M. A. Arbib (Ed.), *Action to Language: via the Mirror Neuron System* (pp. 347-373). Cambridge, UK: Cambridge University Press.

Kemmerer, D., & Eggleston, A. (2010). Nouns and verbs in the brain: Implications of linguistic typology for cognitive neuroscience. *Lingua, 120*(12), 2686-2690.

Kemmerer, D., Gonzalez-Castillo, J., Talavage, T., Petterson, S., & Wiley, C. (2008). Neuroanatomical distribution of five semantic components of verbs: Evidence from fMRI. *Brain & Language, 107*(1), 16-43.

Kim, J. G., & Biederman, I. (2010). Where do objects become scenes? *Cerebral Cortex, 21*(8), 1738-1746.

Kim, M.-S., & Cave, K. R. (1995). Spatial Attention in Visual Search for Features and Feature Conjunctions. *Psychological Science, 6*(6), 376-380.

King, J. A., Burgess, N., Hartley, T., Vargha-Khadem, F., & O'Keefe, J. (2002). Human hippocampus and viewpoint dependence in spatial memory. *Hippocampus, 12*(6), 811-820.

Knops, A., Thirion, B., Hubbard, E. M., Michel, V., & Dehaene, S. (2009). Recruitment of an Area Involved in Eye Movements During Mental Arithmetic. *Science, 324*(5934), 1583-1585.

Knott, A. (2003). Grounding Syntactic Representations in an Architecture for Sensorimotor Control.

Koenigs, M., Barbey, A. K., Postle, B. R., & Grafman, J. (2009). Superior Parietal Cortex Is Critical for the Manipulation of Information in Working Memory. *The Journal of Neuroscience, 29*(47), 14980-14986.

Kosslyn, S. M. (1987). Seeing and imaging in the cerebral hemispheres: A computational approach. *Psychological Review, 94*(2), 148-175.

Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate*. Cambrdige, MA: The MIT press.

Kowler, E., & Anton, S. (1987). Reading twsited text: Implications for the role of saccades. *Vision Research, 27*(1), 45-60.

Kuchinsky, S. E. (2009). *From Seeing to Saying: Perceiving, Planning, Producing.* Ph.D., University of Illinois at Urbana-Champaign, Urbana-Champaign.

Lambon Ralph, M. A., Lowe, C., & Rogers, T. T. (2007). Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model. *Brain, 130*(4), 1127-1137.

Landau, B., & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences, 16*(2), 217-238.

Langacker, R. W. (1987). *Foundations of Cognitive Grammar: Theoretical prerequisites* (Vol. 1). Stanford: Stanford University Press.

Langacker, R. W. (1991). *Foundations of Cognitive Grammar: Descriptive application* (Vol. 2). Stanford: Stanford University Press.

Lashley, K. S. (1951). The Problem of Serial Order in Behavior. In L. A. Jeffress (Ed.), *Cerebral Mechanisms in Behavior* (pp. 112-136). New York: John Wiley & Sons.

Lavie, N. (2005). Distracted and confused?: Selective attention under load. *TRENDS in Cognitive Sciences, 9*(2), 75-82.

Lavie, N., & Cox, S. (1997). On the Efficiency of Visual Selective Attention: Efficient Visual Search Leads to Inefficient Distractor Rejection. *Psychological Science, 8*(5), 395-398.

Lebedev, M. A., Messinger, A., Kralik, J. D., & Wise, S. P. (2004). Representation of Attended Versus Remembered Locations in Prefrontal Cortex. *PLoS Biology, 2*(11), 1919-1935.

Levelt, W. J. M. (1989). *Speaking: From Intention to Articulation*. London: MIT Press.

Levelt, W. J. M., & Meyer, A. S. (2000). Word for word: Multiple lexical access in speech production. *European Journal of Cognitive Psychology, 12*(4), 433-452.

Levine, R. D., & Meurers, D. (2006). Head-Driven Phrase Structure Grammar: Linguistic Approach, Formal Foundations, and Computational Realization. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics, Second Edition*: Oxford: Elsevier.

Levy, R., & Goldman-Rakic, P. S. (2000). Segregation of working memory functions within the dorsolateral prefrontal cortex. *Experimental Brain Research, 133*(1), 23-32.

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research, 25*(1), 93-115.

Li, L.-J., Socher, R., & Fei-Fei, L. (2009). *Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework.* Paper presented at the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*(6657), 279-281.

Lumer, E. D., Friston, K. J., & Rees, G. (1998). Neural Correlates of Perceptual Rivalry in the Human Brain. *Science, 280*(5371), 1930-1934.

Ma, Y., Hu, X., & Wilson, F. A. W. (2011). The egocentric spatial reference frame used in dorsal-lateral prefrontal working memory in primates. *Neuroscience & Biobehavioral Reviews, 36*(1), 26-33.

Ma, Y., Ryou, J.-W., Kim, B.-H., & Wilson, F. A. W. (2004). Spatially directed movement and neuronal activity in freely moving monkey. *Progress in Brain Research, 143*, 513-520.

Ma, Y., Tian, B. P., & Wilson, F. A. W. (2003). Dissociation of egocentric and allocentric spatial processing in prefrontal cortex. *Neuroreport, 14*(13), 1737-1741.

Maess, B., Koelsch, S., Gunter, T. C., & Friederici, A. D. (2001). Musical syntax is processed in Broca's area: an MEG study. *Nature Neuroscience, 4*(5), 540-545.

Mahon, B. Z., & Caramazza, A. (2005). The orchestration of the sensory-motor systems: Clues from Neuropsychology. *Cognitive Neuropsychology, 22*(3-4), 480-494.

Mahon, B. Z., & Caramazza, A. (2008). A critical look at the embodied cognition hypothesis and a new proposal for grounding conceptual content. *Journal of Physiology-Paris, 102*(1-3), 59-70.

Majerus, S., Poncelet, M., van der Linden, M., Albouy, G., Salmon, E., Sterpenich, V., . . . Maquet, P. (2006). The left intraparietal sulcus and verbal short-term memory: focus of attention or serial order? *NeuroImage, 32*(2), 880-891.

Marr, D. (1983). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY: Henry Holt and Co., Inc.

Martin, A., & Chao, L. L. (2001). Semantic memory and the brain: structure and processes. *Current Opinion in Neurobiology, 11*(2), 194-201.

Martin, A., Wiggs, C. L., Ungerleider, L. G., & Haxby, J. V. (1996). Neural correlates of category-specific knowledge. *Nature, 379*(6566), 649-652.

Martin, R. C., Crowther, J. E., Knight, M., Tamborello II, F. P., & Yang, C.-L. (2010). Planning in sentence production: Evidence for the phrase as a default planning scope. *Cognition, 116*(2), 177-192.

Martin, R. C., & Freedman, M. L. (2001). Short-term retention of lexical-semantic representations: Implications for speech production. *Memory, 9*(4-6), 261-280.

Matthei, E. H. (1982). The acquisition of prenominal modifier sequences. *Cognition, 11*(3), 301-332.

McCrea, S. M., Buxbaum, L. J., & Coslett, H. B. (2006). Illusory conjunctions in simultanagnosia: Coarse coding of visual feature location? *Neuropsychologia, 44*(10), 1724-1736.

McEvoy, L. K., Smith, M. E., & Gevins, A. (1998). Dynamic cortical networks of verbal and spatial working memory: effects of memory load and task practice. *Cerebral Cortex, 8*(7), 563-574.

Meyer, A. S. (2004). The Use of Eye Tracking in Studies of Sentence Generation. In J. M. Henderson & F. Ferreira (Eds.), *The Interface of Language, Vision, and Action* (pp. 191-212). New York, NY: Psychology Press.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. M. (1998). Viewing and naming objects: eye movements during noun phrase production. *Cognition, 66*(2), B25-B33.

Meyer, A. S., Van der Meulen, F. F., & Brooks, A. (2004). Eye movements during speech planning: Talking about present and

remembered objects. *Visual Cognition, 11*(5), 553-576.

Michaelis, L. A., & Lambrecht, K. (1996). Toward a Construction-Based Model of Language Function: The Case of Nominal Extraposition. *Language, 72*(2), 215-247.

Miikkulainen, R., & Dyer, M. G. (1991). Naturallanguageprocessing with modularpdpnetworks and distributedlexicon. *Cognitive Science, 15*(3), 343-399.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24*(1), 167-202.

Miller, E. K., & Desimone, R. (1994). Parallel neuronal mechanisms for short-term memory. *Science, 263*(5146), 520-522.

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural Mechanisms of Visual Working Memory in Prefrontal Cortex of the Macaque. *The Journal of Neuroscience, 16*(16), 5154-5167.

Milner, A. D., & Goodale, M. A. (1995). *The Visual Brain in Action*. Oxford: Oxford University Press.

Mishkin, M., & Ungerleider, L. G. (1982). Contribution of striate inputs to the visuospatial functions of parieto-preoccipital cortex in monkeys. *Behavioral and Brain Research, 6*(1), 57-77.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in Neuroscience, 6*(10), 414-417.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How Are Visuospatial Working Memory, Executive Functioning, and Spatial Abilities Related? A Latent-Variable Analysis. *Journal of Experimental Psychology: General, 130*(4), 621-640.

Mokler, A., & Fischer, B. (1999). The recognition and correction of involuntary prosaccades in an antisaccade task. *Experimental Brain Research, 125*(4), 511-516.

Moser, D., Baker, J. M., Sanchez, C. E., Rorden, C., & Fridriksson, J. (2009). Temporal Order Processing of Syllables in the Left Parietal Lobe. *The Journal of Neuroscience, 29*(40), 12568-12573.

Müller, N. G., Bartelt, O. A., Donner, T. H., Villringer, A., & Brandt, S. A. (2003). A Physiological Correlate of the "Zoom Lens" of Visual Attention. *The Journal of Neuroscience, 23*(9), 3561-3565.

Müller, R.-A., & Basho, S. (2004). Are nonlinguistic functions in "Broca's area" prerequisites for language acquisition? FMRI findings from an ontogenetic viewpoint. *Brain & Language, 89*(2), 329-336.

Mullette-Gillman, O. D. A., Cohen, Y. E., & Groh, J. M. (2009). Motor-Related Signals in the Intraparietal Cortex Encode Locations in a Hybrid, rather than Eye-Centered Reference Frame. *Cerebral Cortex, 19*(8), 1761-1775.

Munk, M. H. J., Linden, D. E. J., Muckli, L., Lanfermann, H., Zanella, F. E., Singer, W., & Goebel, R. (2002). Distributed Cortical Systems in Visual Short-term Memory Revealed by Event-related Functional Magnetic Resonance Imaging. *Cerebral Cortex, 12*(8), 866-876.

Murata, A., Gallese, V., Luppino, G., Kaseda, M., & Sakata, H. (2000). Selectivity for the Shape, Size, and Orientation of Objects for Grasping in Neurons of Monkey Parietal Area AIP. *Journal of Neurophysiology, 83*(5), 2580-2601.

Murray, E. A., & Richmond, B. J. (2001). Role of perirhinal cortex in object perception, memory, and associations. *Current Opinion in Neurobiology, 11*(2), 188-193.

Narayanan, S. S. (1997). *Knowledge-based Action Representations for Metaphor and Aspect (KARMA).* Ph. D., University of

California, Berkeley, Berkeley, CA.

Narayanan, S. S. (1999). *Moving right along: a computational model of metaphoric reasoning about events.* Paper presented at the Proceedings of the sixteenth national conference on Artificial intelligence and the eleventh Innovative applications of artificial intelligence conference innovative applications of artificial intelligence, Orlando, FL.

Navalpakkam, V., & Itti, L. (2005). Modeling the influence of task on attention. *Vision Research, 45*(2), 205-231.

Negri, G. A. L., Rumiati, R. I., Zadini, A., Ukmar, M., Mahon, B. Z., & Caramazza, A. (2007). What is the role of motor simulation in action and object recognition? Evidence from apraxia. *Cognitive Neuropsychology, 24*(8), 795-816.

Ninokura, Y., Mushiake, H., & Tanji, J. (2004). Integration of Temporal Order and Object Information in the Monkey Lateral Prefrontal Cortex. *Journal of Neurophysiology, 91*(1), 555-560.

Nissen, M. J. (1985). Accessing Features and Objects; Is Location Special? In M. I. Posner & O. S. Marin (Eds.), *Attention and performance XI* (pp. 205-219). Hillsdale, NJ: Lawrence Erlbaum.

Nobre, A. C., Coull, J. T., Maquet, P., Frith, C. D., Vandenberghe, R., & Mesulam, M. M. (2004). Orienting Attention to Locations in Perceptual Versus Mental Representations. *Journal of Cognitive Neuroscience, 16*(3), 363-373.

Noori, N., & Itti, L. (2011). Visuospatial attention shifts during non-visual mental tasks. *Journal of Vision, 11*(11), article 479.

Nyberg, L., Marklund, P., Persson, J., Cabeza, R., Forkstam, C., Petersson, K. M., & Ingvar, M. (2003). Common prefrontal activations during working memory, episodic memory, and semantic memory. *Neuropsychologia, 41*(3), 371-377.

Nyberg, L., McIntosh, A. R., Cabeza, R., Habib, R., Houle, S., & Tulving, E. (1996). General and specific brain regions involved in encoding and retrieval of events: What, where, and when. *Proceedings of the National Academy of Sciences of the United States of America, 93*(20), 11280-11285.

O'Keefe, J. (1999). Do hippocampal pyramidal cells signal non-spatial as well as spatial information? *Hippocampus, 9*(4), 352-364.

O'Regan, J. K. (1992). Solving the "Real" mysteries of visual perception: the world as an outside memory. *Canadian Journal of Psychology, 46*(3), 461-488.

Oberauer, K. (2002). Access to Information in Working Memory: Exploring the Focus of Attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(3), 411-421.

Offen, S., Gardner, J. L., Schluppeck, D., & Heeger, D. J. (2010). Differential roles for frontal eye fields (FEFs) and intraparietal sulcus (IPS) in visual working memory and visual attention. *Journal of Vision, 10*(11), 1-14.

Oh, S.-H., & Kim, M.-S. (2003). The guidance effect of working memory load on visual search. *Journal of Vision, 3*(9), article 629.

Oliva, A., & Schyns, P. G. (1997). Coarse blobs or fine edges? Evidence that information diagnosticity changes the perception of complex visual stimuli. *Cognitive Psychology, 34*(1), 72-107.

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision, 42*(3), 145-175.

Oppermann, F., Jescheniak, J. D., & Schriefers, H. (2010). Phonological advance planning in sentence production. *Journal of Memory and Language, 63*(4), 526-540.

Osaka, N., Osaka, M., Kondo, H., Morishita, M., Fukuyama, H., & Shibasaki, H. (2003). The neural basis of executive

function in working memory: an fMRI study based on individual differences. *NeuroImage, 21*(2), 623-631.

Osgood, C. E. (1977). Salience and sentencing: some production principles. In S. Rosenberg (Ed.), *Sentence production: Developments in research and theory*. Hillsdale, NJ: Erlbaum.

Owen, A. M., Evans, A. C., & Petrides, M. (1996). Evidence for a Two-Stage Model of Spatial Working Memory Processing within the Lateral Frontal Cortex: A Positron Emission Tomography Study. *Cerebral Cortex, 6*(1), 31-38.

Papafragou, A., Hulbert, J., & Trueswell, J. C. (2008). Does language guide event perception? Evidence from eye movements. *Cognition, 108*(1), 155-184.

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., & Smith, A. D. (2002). Models of Visuospatial and Verbal Memory Across the Adult Life Span. *Psychology and Aging, 17*(2), 299-320.

Pecher, D., Zeelenberg, R., & Barsalou, L. W. (2003). Verifying Properties from Different Modalities for Concepts Produces Switching Costs. *Psychological Science, 14*(2), 119-124.

Peelen, M. V., Fei Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature, 460*(2), 94-97.

Petrides, M. (2000). The role of the mid-dorsolateral prefrontal cortex in working memory. *Experimental Brain Research, 133*(1), 44-54.

Pierrot-Deseilligny, C., Müri, R. M., Rivaud-Pechoux, S., Gaymard, B., & Ploner, C. J. (2002). Cortical control of spatial memory in humans: The visuooculomotor model. *Annals of Neurology, 52*(1), 10-19.

Piñango, M. M. (2006). Understanding the architecture of language: the possible role of neurology. *TRENDS in Cognitive Sciences, 10*(2), 49-51.

Piñango, M. M., & Zurif, E. B. (2001). Semantic Operations in Aphasic Comprehension: Implications for the Cortical Organization of Language. *Brain & Language, 79*(2), 297-308.

Pinker, S. (1989). *Learnability and Cognition: The Acquisition of Argument Structure*. Cambridge, Massachusetts: The MIT Press.

Pinsk, M. A., Doniger, G. M., & Kastner, S. (2004). Push-Pull Mechanism of Selective Attention in Human Extrastriate Cortex. *Journal of Neurophysiology, 92*(1), 622-629.

Poeppel, D., & Hickok, G. (2004). Towards a new functional anatomy of language. *Cognition, 92*(1-2), 1-12.

Pollard, C., & Sag, I. A. (1994). *Head-driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology, 32*(1), 3-25.

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience, 139*(1), 23-38.

Postle, B. R., Idzikowski, C., Sala, S. D., Logie, R. H., & Baddeley, A. D. (2006). The selective disruption of spatial working memory by eye movements. *The Quarterly Journal of Experimental Psychology, 59*(1), 100-120.

Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2*(5), 509-522.

Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences, 1191*(1), 62-88.

Price, C. J., Thierry, G., & Griffiths, T. (2005). Speech-specific auditory processing: where is it? *TRENDS in Cognitive*

Sciences, 9(6), 271-276.

Pulvermüller, F. (2001). Brain reflections of words and their meaning. *TRENDS in Cognitive Sciences, 5*(12), 517-524.

Pulvermüller, F., & Fadiga, L. (2010). Active perception: sensorimotor circuits as a cortical basis for language. *Nature Reviews Neuroscience, 11*(5), 351-360.

Pulvermüller, F., Hauk, O., Nikulin, V. V., & Ilmoniemi, R. J. (2005). Functional links between motor and language systems. *European Journal of Neuroscience, 21*(3), 793-797.

Pulvermüller, F., Mohr, B., & Schleichert, H. (1999). Semantic or lexico-syntactic factors: what determines word-class specific activity in the human brain? *Neuroscience Letters, 275*(2), 81-84.

Pylyshyn, Z. W. (1984). *Computation and cognition: toward a foundation for cognitive science*. Cambridge, MA: MIT Press.

Pylyshyn, Z. W. (2001). Visual indexes, preconceptual objects, and situated vision. *Cognition, 80*(1-2), 127-158.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision, 3*(3), 179-197.

Qi, X.-L., Katsuki, F., Meyer, T., Rawley, J. B., Zhou, X., Douglas, K. L., & Constantinidis, C. (2011). Comparison of neural activity related to working memory in primate dorsolateral prefrontal and posterior parietal cortex. *Frontiers in Systems Neuroscience, 4*, 12.

Rainer, G., Asaad, W. F., & Miller, E. K. (1998). Memory fields of neurons in the primate prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 95*(25), 15008-15013.

Ranganath, C., DeGutis, J., & D'Esposito, M. (2004). Category-specific modulation of inferior temporal activity during working memory encoding and maintenance. *Cognitive Brain Research, 20*(1), 37-45.

Rao, S. C., Rainer, G., & Miller, E. K. (1997). Integration of What and Where in the Primate Prefrontal Cortex. *Science, 276*(5313), 821-824.

Rappaport Hovav, M., & Levin, B. (1998). Building Verb Meanings. In M. Butt & W. Geuder (Eds.), *The projection of arguments: Lexical and syntactic constraints* (pp. 97-134). Stanford, CA: CSLI Publications.

Raymond, J. E., Shapiro, K. L., & Arnell, K. M. (1992). Temporary Suppression of Visual Processing in an RSVP Task: An Attentional Blink? *Journal of Experimental Psychology: Human Perception and Performance, 18*(3), 849-860.

Rensink, R. A. (2000a). The Dynamic Representation of Scenes. *Visual Cognition, 7*(1-3), 17-42.

Rensink, R. A. (2000b). Seeing, sensing, and scrutinizing. *Vision Research, 40*(10-12), 1469-1487.

Ricci, C., & Blundo, C. (1990). Perception of ambiguous figures after focal brain lesions. *Neuropsychologia, 28*(11), 1163-1173.

Richardson, D. C., & Spivey, M. J. (2000). Representation, space and Hollywood Squares: looking at things that aren't there anymore. *Cognition, 76*(3), 269-295.

Rilling, J. K., Glasser, M. F., Preuss, T. M., Ma, X., Zhao, T., Hu, X., & Behrens, T. E. J. (2008). The evolution of the arcuate fasciculus revealed with comparative DTI. *Nature Neuroscience, 11*(4), 426-428.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition, 42*(1-3), 107-142.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and Deterioration of Semantic Memory: A Neuropsychological and Computational Investigation.

Psychological Review, 111(1), 205-235.

Rolls, E. T., Aggelopoulos, N. C., & Zheng, F. (2003). The Receptive Fields of Inferior Temporal Cortex Neurons in Natural Scenes. *The Journal of Neuroscience, 23*(1), 339-348.

Romanski, L. M. (2007). Representation and Integration of Auditory and Visual Stimuli in the Primate Ventral Lateral Prefrontal Cortex. *Cerebral Cortex, 17*(suppl 1), i61-i69.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382-439.

Rumelhart, D. E. (1977). *Introduction to human information processing* New York, NY: Wiley.

Rushworth, M. F. S., Johansen-Berg, H., Göbel, S. M., & Devlin, J. T. (2003). The left parietal and premotor cortices: motor attention and selection. *NeuroImage, 20*(Suppl 1), S89-S100.

Rypma, B., Berger, J. S., & D'Esposito, M. (2002). The Influence of Working-Memory Demand and Subject Performance on Prefrontal Cortical Activity. *Journal of Cognitive Neuroscience, 14*(5), 721-731.

Sala, J. B., & Courtney, S. M. (2007). Binding of What and Where During Working Memory Maintenance. *Cortex, 43*(1), 5-21.

Sala, J. B., Rämä, P., & Courtney, S. M. (2003). Functional topography of a distributed neural system for spatial and nonspatial information maintenance in working memory. *Neuropsychologia, 41*(3), 341-356.

Saur, D., Kreher, B. W., Schnell, S., Kummerer, D., Kellmeyer, P., Vry, M.-S., . . . Weiller, C. (2008). Ventral and dorsal pathways for language. *Proceedings of the National Academy of Sciences of the United States of America, 105*(46), 18035-18040.

Sawaguchi, T., & Iba, M. (2001). Prefrontal Cortical Representation of Visuospatial Working Memory in Monkeys Examined by Local Inactivation With Muscimol. *Journal of Neurophysiology, 86*(4), 2041-2053.

Saygin, A. P., Dick, F., Wilson, S. W., Dronkers, N. F., & Bates, E. (2003). Neural resources for processing language and environmental sounds: Evidence from aphasia. *Brain, 126*(4), 928-945.

Schnur, T. T. (2011). Phonological Planning during Sentence Production: Beyond the Verb. *Frontiers in Psychology, v*.

Schnur, T. T., Costa, A., & Caramazza, A. (2006). Planning at the Phonological Level during Sentence Production. *Journal of Psycholinguistic Research, 35*(2), 198-213.

Scholl, B. J. (2001). Objects and attention: the state of the art. *Cognition, 80*(1-2), 1-46.

Scholl, B. J., & Pylyshyn, Z. W. (1999). TrackingMultipleItems Through Occlusion: Clues to VisualObjecthood. *Cognitive Psychology, 38*(2), 259-290.

Scholl, B. J., Pylyshyn, Z. W., & Feldman, J. (2001). What is a visual object? Evidence from target merging in multiple object tracking. *Cognition, 80*(1-2), 159-177.

Scholl, B. J., Pylyshyn, Z. W., & Franconeri, S. L. (2004). *The relationship between property-encoding and object-based attention: Evidence from multiple object tracking*. Manuscript submitted for publication.

Schwartz, S., Vuilleumier, P., Hutton, C., Maravita, A., Dolan, R. J., & Driver, J. (2005). Attentional Load and Sensory Competition in Human Vision: Modulation of fMRI Responses by Load at Fixation during Task-irrelevant Stimulation in the Peripheral Visual Field. *Cerebral Cortex, 15*(6), 770-786.

Selkirk, E. O. (1984). *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge: MIT Press.

Shastri, L. (2002). Episodic memory and cortico-hippocampal interactions. *TRENDS in Cognitive Sciences, 6*(4), 162-168.

Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences, 16*(3), 417-451.

Simmons, W. K., & Barsalou, L. W. (2003). The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cognitive Neuropsychology, 20*(3-6), 451-486.

Simons, D. J., & Rensink, R. A. (2005). Change blindness: past, present, and future. *TRENDS in Cognitive Sciences, 9*(1), 16-20.

Smith, E. E., & Jonides, J. (1999). Storage and Executive Processes in the Frontal Lobes. *Science, 283*(5408), 1657-1661.

Smith, E. E., Jonides, J., Marshuetz, C., & Koeppe, R. A. (1998). Components of verbal working memory: Evidence from neuroimaging. *Proceedings of the National Academy of Sciences of the United States of America, 95*(3), 876-882.

Smith, E. E., & Medin, D. L. (1981). *Categories and Concepts*. Cambridge, MA: Harvard University Press.

Smyth, M. M., & Scholey, K. A. (1994). Interference in immediate spatial memory. *Memory and Cognition, 22*(1), 1-13.

Snyder, L. H., Batista, A. P., & Andersen, R. A. (1997). Coding of intention in the posterior parietal cortex. *Nature, 386*(6621), 167-170.

Soto, D., Humphreys, G. W., & Heinke, D. (2006). Working memory can guide pop-out search. *Vision Research, 46*(6-7), 1010-1018.

Sowa, J. F. (2006). Semantic Networks. In L. Nadel (Ed.), *Encyclopedia of Cognitive Science*: John Wiley & Sons, Ltd.

Spinks, J. A., Zhang, J. X., Fox, P. T., Gao, J.-H., & Tan, L. H. (2004). More workload on the central executive of working memory, less attention capture by novel visual distractors: evidence from an fMRI study. *NeuroImage, 23*(2), 517-524.

Spivey, M. J., Richardson, D. C., & Fitneva, S. A. (2004). Thinking outside the brain: Spatial indices to visual and linguistic information. In J. M. Henderson & F. Ferreira (Eds.), *The Interface of Language, Vision, and Action* (pp. 161-190). New York, NY: Psychology Press.

Steels, L., & De Beule, J. (2006a). *Unify and Merge in Fluid Construction Grammars.* Paper presented at the The Third International Workshop on the Emergence and Evolution of Linguistic Communication, Rome, Italy.

Steels, L., & De Beule, J. (2006b). *A (very) Brief Introduction to Fluid Construction Grammar.* Paper presented at the Proceedings of the 3rd Workshop on Scalable Natural Language Understanding, New York City.

Stowe, L. A., Withaar, R. G., Wijers, A. A., Broere, C. A. J., & Paans, A. M. J. (2002). Encoding and Storage in Working Memory during Sentence Comprehension. In P. Merlo & S. Stevenson (Eds.), *The Lexical Basis of Sentence Processing: Formal, computational and experimental issues* (pp. 181-206). Amsterdam: John Benjamins Publishing Company.

Stromswold, K., Caplan, D., Alpert, N., & Rauch, S. (1996). Localization of Syntactic Comprehension by Positron Emission Tomography. *Brain & Language, 52*(3), 452-473.

Stuss, D. T. (2011). Functions of the Frontal Lobes: Relation to Executive Functions. *Journal of the International Neuropsychological Society, 17*(5), 759-765.

Sudderth, E. B., Torralba, A., Freeman, W. T., & Willsky, A. S. (2005). *Learning Hierarchical Models of Scenes, Objects, and Parts.* Paper presented at the 2005 IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION, Beijing.

Suganuma, M., & Yokosawa, K. (2006). Grouping and trajectory storage in multiple object tracking: impairments due to common item motions. *Perception, 35*(4), 483-495.

Talmy, L. (2000). *Toward a Cognitive Semantics* (Vol. 1). Cambridge, MA: MIT Press.

Tettamanti, M., Buccino, G., Saccuman, M. C., Gallese, V., Danna, M., Scifo, P., & Fazio, F. (2005). Listening to Action-related Sentences Activates Fronto-parietal Motor Circuits. *Journal of Cognitive Neuroscience, 17*(2), 273-281.

Tettamanti, M., Rotondi, I., Perani, D., Scotti, G., Fazio, F., Cappa, S. F., & Moro, A. (2009). Syntax without language: Neurobiological evidence for cross-domain syntactic computations. *Cortex, 45*(7), 825-838.

Thiebaut de Schotten, M., Urbanski, M., Duffau, H., Volle, E., Levy, R., Dubois, B., & Bartolomeo, P. (2005). Direct Evidence for a Parietal-Frontal Pathway Subserving Spatial Awareness in Humans. *Science, 309*(5744), 2226-2228.

Todd, J. J., & Marois, R. (2004). Capacity limit of visual short-term memory in human posterior parietal cortex. *Nature, 428*(6984), 751-754.

Tomlin, R. S. (1997). Mapping conceptual representations into linguistic representations: the role of attention in grammar. In J. Nuyts & E. Pederson (Eds.), *Language and conceptualization* (pp. 162-189). Cambridge, UK: Cambridge University Press.

Tranel, D., Kemmerer, D., Adolphs, R., Damasio, H., & Damasio, A. R. (2003). Neural correlates of conceptual knowledge for actions. *Cognitive Neuropsychology, 20*(3-6), 409-432.

Tremblay, S., Saint-Aubin, J., & Jalbert, A. (2006). Rehearsal in serial memory for visual-spatial information: Evidence from eye movements. *Psychonomic Bulletin & Review, 13*(3), 452-457.

Treue, S. (2003). Climbing the cortical ladder from sensation to perception. *TRENDS in Cognitive Sciences, 7*(11), 469-471.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological Review, 101*(1), 80-102.

Tse, P. U., Sheinberg, D. L., & Logothetis, N. K. (2003). Attentional enhancement opposite a peripheral flash revealed using change blindness. *Pyschological Science, 14*(2), 91-99.

Tuholski, S. W., Engle, R. W., & Baylis, G. C. (2001). Individual differences in working memory capacity and enumeration. *Memory and Cognition, 29*(3), 484-492.

Turatto, M., Sandrini, M., & Miniussi, C. (2004). The role of the right dorsolateral prefrontal cortex in visual change awareness. *Neuroreport, 15*(16), 2549-2552.

Tyler, L. K., & Moss, H. E. (2001). Towards a distributed account of conceptual knowledge. *Trends in Cognitive Sciences, 5*(6), 244-252.

Ullman, M. T. (2001). A neurocognitive perspective on language: the declarative/procedural model. *Nature Reviews Neuroscience, 2*(10), 717-726.

Ullman, M. T. (2004). Contributions of memory circuits to language: the declarative/procedural model. *Cognition, 92*(1-2), 231-270.

Ullman, M. T., Pancheva, R., Love, T., Yee, E., & Swinney, D. (2005). Neuralcorrelates of lexicon and grammar: Evidence

from the production, reading, and judgment of inflection in aphasia. *Brain and Language, 93*(2), 185-238.

Ungerleider, L. G., & Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Current Opinion in Neurobiology, 4*(2), 157-165.

van der Meulen, F. F. (2001). *Moving eyes and naming objects.* Ph.D., Nijmegen University, Nijmegen.

van der Meulen, F. F. (2003). Coordination of eye gaze and speech in sentence production. In H. Hartl & H. Tappe (Eds.), *Mediating between concepts and grammar* (pp. 38-64). Berlin: Mouton de Gruyter.

van der Meulen, F. F., Meyer, A. S., & Levelt, W. J. M. (2001). Eye movements during the production of nouns and pronouns. *Memory and Cognition, 29*(3), 512-521.

van der Velde, F., van der Voort van der Kleij, G. T., & Kamps, M. d. (2004). Lack of combinatorial productivity in language processing with simple recurrent networks. *Connection Science, 16*(1), 21-46.

Vargha-Khadem, F., Gadian, D. G., & Mishkin, M. (2001). Dissociations in cognitive memory: the syndrome of developmental amnesia. *Philosophical Transactions of the Royal Society of Biological Sciences, 356*(1413), 1435-1440.

Verhagen, A. (2009). The conception of constructions as complex signs: Emergence of structure and reduction to usage. *Constructions and Frames, 1*(1), 119-152.

Verhagen, A. (2010). What Do You Think is the Proper Place of Recursion? Conceptual and Empirical Issues. In H. van der Hulst (Ed.), *Recursion and Human Language* (pp. 94-110). Berlin/New York: Mouton de Gruyter.

Vigliocco, G., Vinson, D. P., Druks, J., Barber, H., & Cappa, S. F. (2010). Nouns and verbs in the brain: A review of behavioural, electrophysiological, neuropsychological and imaging studies. *Neuroscience & Biobehavioral Reviews, 35*(3), 407-426.

Vosse, T., & Kempen, G. (2000). Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar. *Cognition, 75*(2), 105-143.

Vosse, T., & Kempen, G. (2009). The Unification Space implemented as a localist neural net: predictions and error-tolerance in a constraint-based parser. *Cognitive Neurodynamics, 3*(4), 331-346.

Wagner, A. D., Maril, A., Bjork, R. A., & Schacter, D. L. (2001). Prefrontal Contributions to Executive Control: fMRI Evidence for Functional Distinctions within Lateral Prefrontal Cortex. *NeuroImage, 14*(6), 1337-1347.

Wagner, A. D., Shannon, B. J., Kahn, I., & Buckner, R. L. (2005). Parietal lobe contributions to episodic memory retrieval. *TRENDS in Cognitive Sciences, 9*(9), 445-453.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain, 107*(3), 829-853.

Webb, A., Knott, A., & MacAskill, M. R. (2010). Eye movements during transitive action observation have sequential structure. *Acta Psychologica, 133*(1), 51-56.

White, I. M., & Wise, S. P. (1999). Rule-dependent neuronal activity in the prefrontal cortex. *Experimental Brain Research, 126*(3), 315-335.

Williams, Z. M., Elfar, J. C., Eskandar, E. N., Toth, L. J., & Assad, J. A. (2003). Parietal activity and the perceived direction of ambiguous apparent motion. *Nature Neuroscience, 6*(6), 616-623.

Wilson, F. A. W., Scalaidhe, S. P. O., & Goldman-Rakic, P. S. (1993). Dissociation of object and spatial processing domains

in primate prefrontal cortex. *Science, 260*(5116), 1955-1958.

Wilson, S. M., Galantucci, S., Tartaglia, M. C., Rising, K., Patterson, D. K., Henry, M. L., . . . Gorno-Tempini, M. L. (2011). Syntactic Processing Depends on Dorsal Language Tracts. *Neuron, 72*(2), 397-403.

Wright, R. D., & Ward, L. M. (2008). *Orienting of attention*. New York, NY: Oxford University Press.

Wu, D. H., Morganti, A., & Chatterjee, A. (2008). Neural substrates of processing path and manner information of a moving event. *Neuropsychologia, 46*(2), 704-713.

Xu, F., & Carey, S. (1996). Infants' Metaphysics: The Case of Numerical Identity. *Cognitive Psychology, 30*(2), 111-153.

Yantis, S. (1992). Multielement visual tracking: attention and perceptual organization. *Cognitive Psychology, 24*(3), 295-340.

Ye, Z., & Zhou, X. (2009). Executive control in language processing. *Neuroscience & Biobehavioral Reviews, 33*(8), 1168-1177.

Yi, D.-J., Woodman, G. F., Widders, D., Marois, R., & Chun, M. M. (2004). Neural fate of ignored stimuli: dissociable effects of perceptual and working memory load. *Nature Neuroscience, 7*(9), 992-996.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin, 127*(1), 3-21.

Zimmer, H. D., Speiser, H. R., & Seidler, B. (2003). Spatio-temporal working-memory and short-term object-location tasks use different memory mechanisms. *Acta Psychologica, 114*(1), 41-65.