A Model of Language Processing and Spatial Reasoning Using Skill Acquisition to Situate Action

Scott A. Douglass (scott.douglass@mesa.afmc.af.mil)
Air Force Research Laboratory, Warfighter Readiness Research Division
6030 S. Kent St, Mesa, AZ 85212 USA

John R. Anderson (ja@cmu.edu)
Department of Psychology, Carnegie Mellon University
5000 Forbes Avenue, Pittsburgh, PA 15213 USA

Abstract

This paper describes empirical and modeling efforts that show it is possible to specify and empirically validate a model of situated action that utilizes symbolic representations and rule-based information processes. The discussed work shows that situated actions based on active perception (Bajcsy, 1988) utilizing learned visual routines (Ullman, 1983) can be modeled using symbolic representations and rules in ACT-R. In addition to demonstrating that situated action can be symbolically modeled, the ACT-R model explains the behavioral data obtained in the series of described experiments.

Keywords: Situated Action; Spatial Reasoning; Computational Model; ACT-R Cognitive Architecture

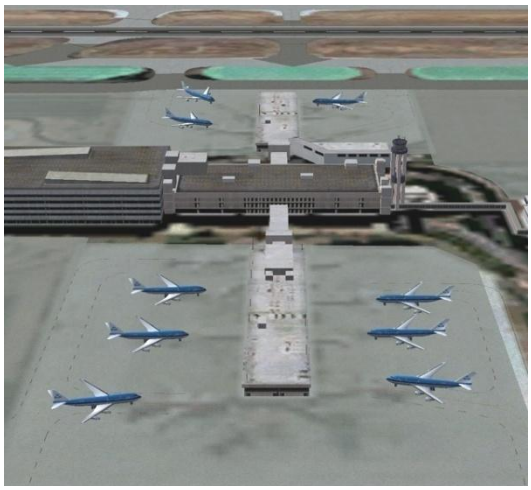
Introduction

The empirical and modeling efforts described in this paper aim to show that the interpretation of a referring expression in a visual context can be achieved through a process combining contextually-situated active vision with simple decision making. The first of these processes represents a rational control of attention utilizing visual routines and behavioral programs that critically depend on unmediated deictic references to referents in the context. The second of these processes represents a reasoned effort to identify the described object. This paper demonstrates how these processes can be plausibly and efficiently modeled in ACT-R (Anderson, Bothell, Byrne, Douglass, Lebiere, & Quin, 2004).

While demonstrating how situated action can be modeled in ACT-R, this paper claims that: (a) skill acquisition can be conceived of as a transition from abstract deliberative behavior to situated reactive behavior; and (b) deictic references explain how perceptually grounded references can incorporate aspects of context into productive variablized patterns of behavior. The first claim is supported by evidence that the ACT-R model of referring expression interpretation matches the performance of human participants by acquiring situated procedural skills through production compilation. These acquired situated procedural skills transition the ACT-R model from the deliberative application of primitive visual routines represented as sets of context-independent productions to the reactive application of composed visual routines represented as sets of context-dependent productions. The second claim is supported by a demonstration that the architecture of ACT-R allows a situated model of referring expression interpretation to index, shift attention to, encode, and incorporate information from context into a variablized behavioral program that spatially reasons.

Description of Experiments

To investigate how people comprehend and act upon referring expressions in visual contexts, two experiments based on a referring expression interpretation task were carried out within the visual world paradigm (Eberhard, Spivey-Knowlton, Sedivy, & Tanenhaus, 1995).



"In group of **three** on **leftmost** ramp, the **middle** plane."
REC1 REC2 REC3

Figure 1: Example trial from the discussed experiments. An example referring expression is included. The labeled words highlighted in red are critical ***R***eferring ***E***xpression ***C***onstraint (REC) terms contained in the expression.

The referring expression interpretation task used in both experiments was based on the "guessing" language game (Steels, 2001). During the experiments, participants attended to referring expressions presented to them aurally through a

text-to-speech (TTS) system and attempted to identify the described referent in a visual context. Figure 1 shows an example visual context.

During the task, participants were asked to listen to each referring expression carefully, decide which airplane was being described, and then indicate their choice with a mouse click. Identified airplanes were highlighted with a blue circle. After three seconds, the airplane actually being described by each referring expression was highlighted in red.

During the experiments, error, latency, and eye movement data was acquired from participants as they interpreted referring expressions.

Materials and Equipment

Custom software presented trials on a 17 inch flat-panel display. The experiment delivery computer used the "Swift" speech synthesis system (developed by Cepstral Software, Pittsburgh) to present referring expressions through a set of multi-media speakers. Experiment delivery software: (a) displayed individual trials to participants, (b) remotely controlled the eye-tracking system described below, and (c) recorded all task performance data.

A SR Research EyelinkII high-speed pupil/corneal reflection eye tracking system was used to track participant eye movements during the experiments. Records of participant actions and word onset/offset times produced by the Swift TTS engine were incorporated into eye movement record files produced by the EyelinkII.

Stimuli

Three forms of referring expressions were presented to participants in the experiments. All three forms consisted of 10 total words, three of which were critical referring expression constraint (REC) terms. Two of the three constraint terms informed the participants how they should limit their visual attention to sub-sets of the candidate objects in the visual contexts. Terms of this sort can be considered context constraints (CC). One of the three constraint terms informed the participants which candidate in a sub-context was the actual referent. Terms of this sort can be considered identity constraints (IC).

In the context of a trial corresponding to Figure 1 above for example, the REC1 and REC2 terms "three" and "leftmost" informed participants that they were to limit their attention to the leftmost group consisting of three airplanes. REC1 and REC2 therefore were the two CCs. Only after limiting their attention to a sub-context could participants utilize the REC3 term "middle" to determine the actual referent being described in the expression. REC3 therefore was the IC.

Table 1 below summarizes the three forms of referring expression presented to participants. Constraint terms were always selected so that the target referent was uniquely described by the constructed referring expression; constraint terms never conflicted with the properties of a visual context. Referring expressions that quickly disambiguated the identity of the referent contained additional redundant constraint terms.

The point at which referential ambiguity was resolved in each referring expression was a function of: (a) how airplanes were configured in the visual contexts, (b) the direction from which the visual contexts were viewed, (c) the view quadrant in which the actual referent was situated, and (d) the format of the referring expression. In any given trial, the point of disambiguation (POD) was REC1, REC2, or REC3. The referring expression in Figure 1 for example disambiguates the referent at REC3 given the visual context. A trial based on the expression and visual context shown in the figure would have a POD at the third constrain term (REC3). A CC-First referring expression starting, "In group of one…" would disambiguate the referent at REC1. A IC-First-O1referring expression starting, "The closest plane in group of two…" would disambiguate the referent at REC2.

In both experiments, trials were presented in 6 blocks of 32 trials. The eye tracker was re-calibrated before the start of each block of trials.

Table 1: The fixed forms of referring expressions used in Experiments.

	REC1	REC2	REC3
Identity Constraint First Order 1 (IC-First-O1)	middle, only farthest, closest leftmost, rightmost	one two three	farthest, closest leftmost, rightmost
Identity Constraint First Order 2 (IC-First-O2)	middle, only farthest, closest leftmost, rightmost	farthest, closest leftmost, rightmost	one two three
Context Constraint First (CC-First)	one two three	farthest, closest leftmost, rightmost	middle, only farthest, closest leftmost, rightmost

The templates for the three forms of referring expression presented during the experiments are:

IC-First-O1: "The <REC1> plane in group of <REC2> on <REC3> ramp."
IC-First-O2: "The <REC1> plane on <REC2> ramp in group of <REC3>.
CC-First: "In group of <REC1> on <REC2> ramp, the <REC3> plane."

Results

Response Times Figures 2 and 3 summarize the latency data produced by participants in the experiments. The average presentation times of REC2 and REC3 are indicated in the figures with labels near the y axis. Confidence intervals (95%) in the figures show that participants were sensitive to the POD in the referring expressions and additionally learned that visual contexts occasionally allowed them to identify the referent before the expression was completely presented.

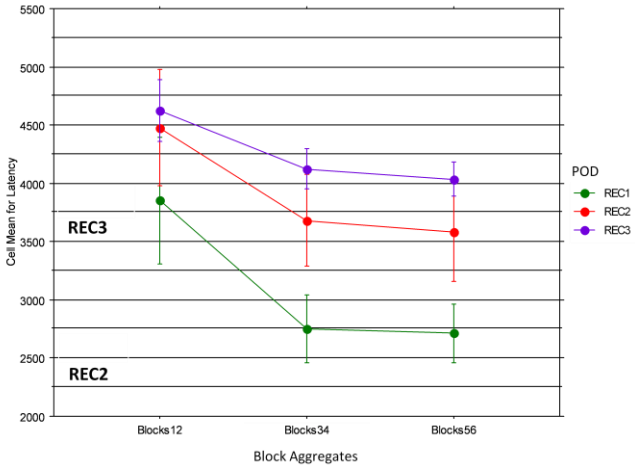


Figure 2: Block Aggregate x POD interaction in Experiment 1. In the experiment, participants interpreted IC-First-O1 and CC-First referring expressions.

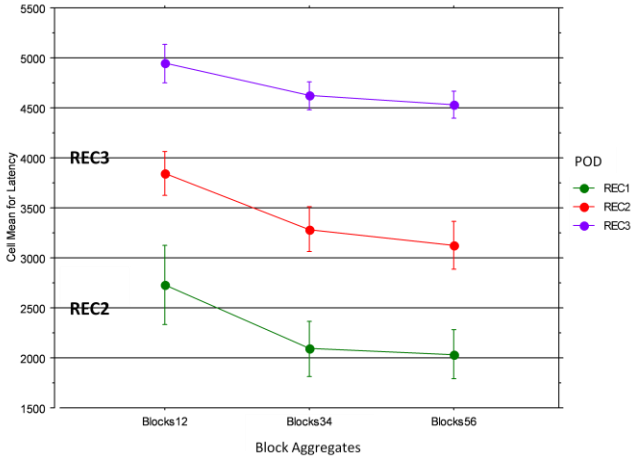


Figure 3: Block Aggregate x POD interaction in Experiment 2. During the experiment, participants interpreted only IC-First-O2 expressions.

In Experiment 1 (Figure 2), during the 1st trial block aggregate, participants generally waited until the presentation of REC3 before acting. By the 2nd and 3rd aggregate blocks, participants routinely responded just after they had attended to the disambiguating REC term. Participants encountering only IC-First-O2 expressions in Experiment 2 (Figure 3) learned to identify referents approximately 500ms after the POD.

In both experiments, participants learned to identify referents shortly after comprehending the POD. Participant response times demonstrate that language processing in the task is both incremental and context dependent.

Eye Movements Eberhard, Spivey-Knowlton, Sedivy, and Tanenhaus (1995) use fine-grained analyses of fixation proportions over time to demonstrate that eye movements are time-locked to language use. They claim that such analyses provide a reliable and fine-grained "profile" of the underlying real-time language processing. They concluded that visual context influences language processing: (a) by providing property clues that help subjects avoid ambiguity, and (b) by providing a "grounding" domain.

To determine the extents to which incremental language processing, systematic ambiguity resolution, and domain grounding influenced the interpretation of referring expressions in Experiments 1 & 2, eye movement analyses similar to those discussed by Eberhard, Spivey-Knowlton, Sedivy, and Tanenhaus (1995) were conducted. To aggregate and compare gaze data across trials and participants, recorded fixations were attributed to one of five POD-sensitive regions of interest (ROI). To determine these ROI within the scope of each trial, the spatial regions corresponding to each of the 4 visible ramps were assigned one of the following ROI labels:

- First RE Constraint (REC1): Regions assigned this label during analysis contained airplanes or groups that became irrelevant as soon as the 1st REC was comprehended. In the time series plots that follow, fixations to regions assigned this label contribute to yellow lines.
- Second RE Constraint (REC2): Regions assigned this label contained airplanes or groups that became irrelevant as soon as the 2nd REC was comprehended. Fixations to regions assigned this label contribute to red lines in the time series plots that follow.
- Third RE Constraint (REC3): Regions assigned this label contained airplanes or groups that became irrelevant as soon as the 3rd REC was comprehended. Fixations to regions assigned this label contribute to purple lines in the time series plots that follow.
- Referring Expression (RE): The region assigned this label met all the constraints conveyed in the RE. In the time series plots that follow, fixations to regions assigned this label contribute to green lines.
- Middle of Nowhere (MON): Fixations not attributable to any of the above ROI were attributed to this ROI. In the time series plots that follow, fixations to regions assigned this label contribute to blue lines.

Figure 4 below presents the time-courses of fixation proportions produced by participants in a particular visual context.

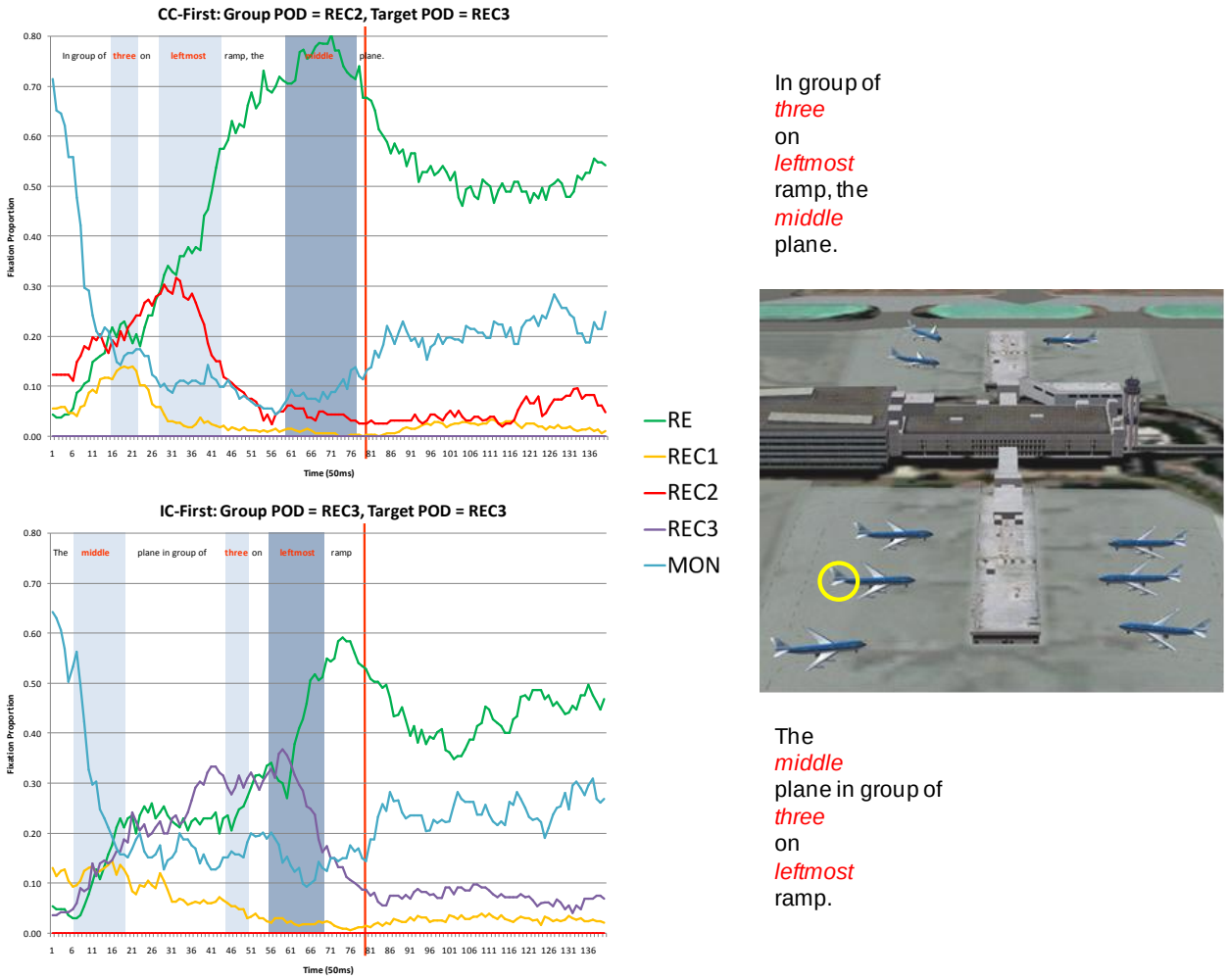


Figure 4: Detailed record of fixation positions during the processing of two referring expressions describing the same referent in different ways. The referent being described is highlighted with a yellow ring.

The graph in the top part of the figure presents the detailed time-course of 10 participants interpreting referring expressions of the CC-First form during blocks 1&2. The graph in the bottom part presents the detailed time-course of 10 participants interpreting referring expressions of the IC-First form during blocks 1&2. Examples of the two different forms of referring expression are provided above and below an example visual context. Intervals during which RECs were aurally presented to participants are indicated with color bands on the graphs. The darker of these onset/offset bands represents the REC that disambiguated the actual referent. The average response time—the point in time at which participants identified the referent they thought was described in the referring expression—is displayed as a solid red vertical line.

The inability of the IC-First referring expression to disambiguate the target's group at REC2 is clearly evident in the lower of the two graphs. In the illustrated visual context, two *slightly* different referring expressions: (a) describe the same referent; (b) differently "define" efficient use of visual attention; and (c) lead to significantly different eye movements.

Discussion

Study of the time course of language processing in the experiments found evidence that the language processes and decision making activities underlying the experimental task were tightly interleaved with active visual exploration.

Fixation proportion time series revealed that participants were able to comprehend RECs approximately 300ms after their onset and use the information they convey to intelligently explore visual contexts. The speed with which participants were able to incorporate constraint knowledge into their exploration of visual context is clear evidence that words are incrementally assigned meaning with situated actions. Fixation proportion time series also revealed that

changes to the presentation order of RECs (different RE-Formats) significantly altered attention shift patterns. Changes in the use of attention resulting from different RE-Formats is clear evidence that the processes underlying context inspection represent a rational control of attention utilizing visual routines.

Description of ACT-R Model

Conceptual Basis of Model

The design of the ACT-R model of situated referent disambiguation is based on three assumptions:

1 Memory and process limitations discourage the representation of un-interpreted information and multiple expression interpretations. The model therefore incrementally interprets referring expressions and develops a single representation of each expression's meaning.

2 Vision is not passive and static; instead it is active, dynamic, probing, exploratory, and most importantly guided by a complex interleaving of visual processing and context assessment (Bajcsy, 1988). The model therefore employs expression and context sensitive productions to intelligently explore visual contexts.

3 In order to reduce the computational complexity of visual processing, the visual system employs composable primitive operations during visual recognition (Ullman, 1983). The model therefore uses ACT-R's production compilation mechanism to learn rules enabling more sophisticated visual routines.

Main Features of Model

The model consists of two relatively independent threads that enable incremental linguistic processing and active vision. The threads can be summarized as follows:

- The Language Processing (LP) thread: Authors such as Garrod & Sanford (1995) have pointed out the importance of incremental semantic analysis when cognitive or environmental constraints are present. Productions constituting the LP thread incrementally construct and refine representations of referring expressions. Productions in the thread enable the model to incrementally; (a) attend to heard words, (b) comprehend heard words, and (c) integrate the meaning of heard and expected words into expression representations.
- The Visual/Spatial Processing (VSP) thread: Authors such as Ullman (1983), Ballard et al. (1997) and Hayhoe (2000) have pointed out the importance of composable active visual operations when visual contexts must be quickly perceived and conceptualized. Productions constituting the VSP thread use composable visual operations that actively perceive and conceptualize visual contexts. The thread allows the model to "soft-assemble" efficient explorations of visual contexts. Through the use of deictic/indexical references, the visual operations also ground objects in situations into symbolic representations and rule-based processes in ACT-R.

The LP and VSP threads are not explicitly interleaved in the model. During early trials, productions in both threads opportunistically fire. Dynamic properties of the environment and the demands of the situation typically prevent the threads from conflicting. For example, when unattended words are heard, initial production utilities lead to the model delaying on-going active vision until the word has been attended to. As the model acquires new procedural skills through production composition, productions in the VSP and LP threads are sometimes combined. When production composition integrates productions from the VSP and LP threads in the context of successful action, the model learns how to "safely" interleave language processing with active vision.

Performance Evaluation of Model

Figure 5 compares the average response times of the model and participants in Experiment 2. Performance has been broken down by block aggregates and POD. The overall correlation and mean deviation between model and participant performance are 0.971 and 238.5 ms respectively. The figure shows the model performance was quite similar to participant performance.

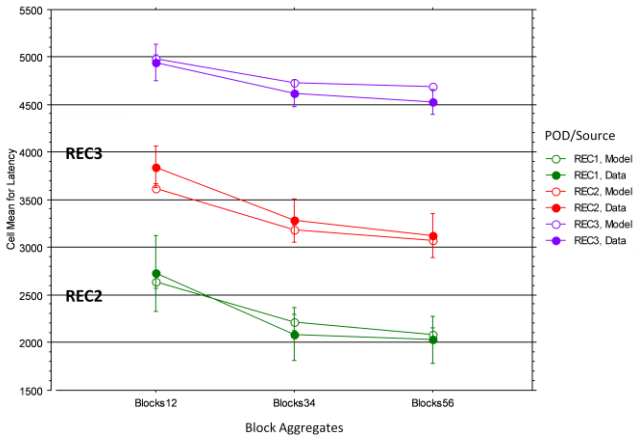


Figure 5: Average response times of the model and Experiment 2 participants.

Latency-based performance comparisons on their own cannot validate the claim that the model performs the reference disambiguation task using the same situated cognitive processes as participants. Since the model ultimately aims to demonstrate that situated action can be modeled in ACT-R, it is critical that the incremental processes of the model be compared to the incremental processes of participants. As the model runs through trials in the experiments, the vision module in ACT-R generates synthetic attention and gaze shifts. These predicted events can be compared to actual participant eye movements.

Figure 6 compares model-derived and participant-derived fixation time series. The model's active visual exploration leaves an attention shift signature similar to that of the participants. The similarities in the signatures suggest that the underlying real-time vision and language processes of the model are similar to those employed by participants.

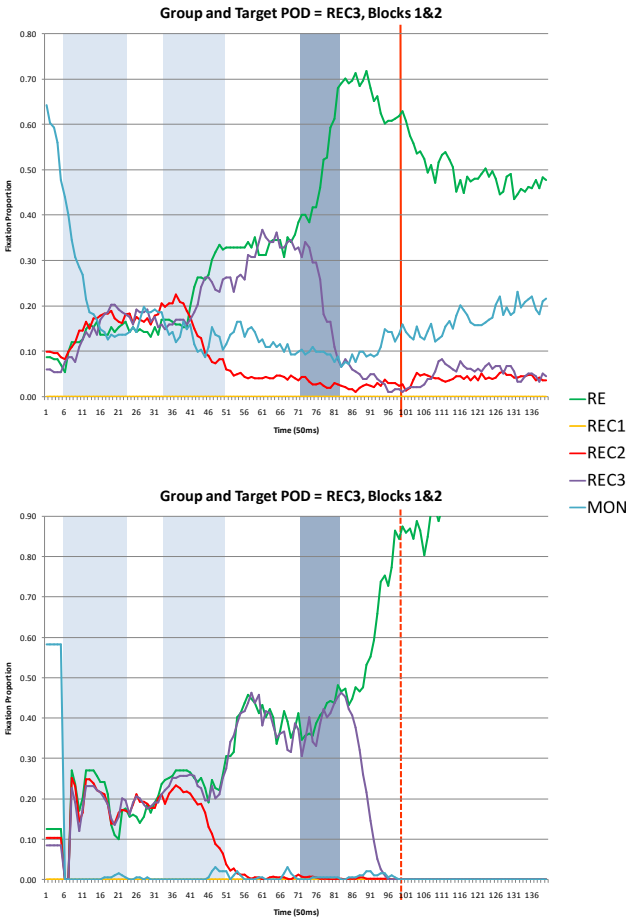


Figure 6: Time series comparison between Experiment 2 participants and the model when POD was REC3.

Table 2 lists aggregate time series correlations between participant and model fixations in Experiments 1 & 2. For brevity, only time series correlations from the first and last block aggregates are included.

Table 2: Aggregate fixation proportion correlations.

		Experiment 1				Experiment 2		
		IC-First		CC-First		IC-First-O2		RE-Format
		1&2	5&6	1&2	5&6	1&2	5&6	Blocks
POD	REC1	0.915	0.946	0.941	0.940	0.916	0.869	
	REC2	0.909	0.926	0.937	0.906	0.948	0.916	
	REC3	0.896	0.899			0.870	0.835	

The correlations show a strong correspondence between the gaze shifts of participants and the model. The correlations are compelling evidence that the model performs the reference disambiguation task using the same situated cognitive processes as participants.

Summary

This paper discussed experimental and modeling efforts that show it's possible to specify a plausible model of referring expression interpretation (situated action) using symbolic representations and rule-based information processes. The discussed model is sensitive to the format of referring expressions, properties of visual contexts, and relationships between constraint terms and the ambiguity properties of the visual contexts in which they distinguish referents. As it learns to anticipate up-coming constraint terms, the model reorganizes its knowledge about expressions and learns to identify referents before comprehending complete referring expressions. The model acquires procedural skills through production compilation. The adoption of these new skills transitions the ACT-R model from the deliberative use of primitive visual routines (context-independent productions) to the reactive use of composed visual routines (context-dependent productions).

Acknowledgements

This research was sponsored by the National Aeronautical and Space Administration (NASA) under contract number NASA-NNA04CLIIA to John R. Anderson at Carnegie Mellon University.

References

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S. A., Lebiere, C., & Quin, Y. (2004). An integrated theory of mind. *Psychological Review , 111*, 1036-1060.

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE , 76* (8), 996-1005.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., & Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences , 20*, 723-767.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., & Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *Journal of Psycholinguistic Research , 24* (6), 409-437.

Garrod, S., & Sanford, A. (1995). Incrementality in discourse understanding. (D. Milward, & P. Sturt, Eds.) *Incremental Interpretation , 11*, pp. 99-122.

Hayhoe, M. (2000). Vision using routines: A functional vision account. *Visual Cognition , 7* (1), 43-61.

Steels, L. (2001). Language games for autonomous robots. *IEEE Intelligent Systems , 16* (5), 16-22.

Ullman, S. (1983). *Visual routines.* A.I. Memo (AIM-732), 1-65, Artificial Intelligence Laboratory: Massachusetts Institute of Technology.