Linguistic Descriptions of Visual Event Perceptions

Anthony B. Maddox
James Pustejovsky

Department of Computer Science
Brandeis University
Waltham, MA 02254
617-736-2700
tony@brandeis.csnet-relay

jamesp@brandeis.csnet-relay

Abstract

In this paper we address the problem of constructing a computational device that is able to describe in natural language its own conceptualization of visual input. This addresses the basic issues of event perception from raw data, as well as what connnection a language with a limited vocabulary has to this event construction. We outline a model of how the perceptual primitives in a system act to both constrain the possible conceptualizations and naturally limit the language used to describe events.

Topic: Visual Perception, Natural Language, Lexical Semantics

1. Introduction

In order for an artificially intelligent system to interact with humans, it is desirable that it be able to communicate with them. Characterizing this interaction will, in part, require considering the impact of language and perception on the communication process. This paper will address the use of language in describing visually perceived events. The focus will be on a theoretical but practical description of the interface between a limited vocabulary linguistic system which supports both tense and aspect and a perceptual representation for visual events. Two major issues discussed are visuo-linguistic temporal granularity and the effect of the interaction between "hard-wired" and learned *focus of attention* on event conceptualization. The paper begins with a discussion of vision-language research and the problems associated with integrating vision and language. In section 3, we present our linguistic and visual concept structures. Section 4 follows with a description of the visuo-linguistic interface illustrated by 4 examples. Section 6 concludes the paper with a summary and directions for future research.

2. Language-Vision Research

There has been little research concerning the interface of visual and linguistic processes. One reason for this is that they each currently appear to involve very different and difficult processes. Considerable energy has been focused on low-level or early vision. Marr's [17] primal sketch includes several low-level primitives from which scenes can be constructed. Many others have developed formalisms that relate low-level visual information to the analysis of polyhedral scenes in the blocks world [1]. Their work points out the difficulty in analyzing even the most simple scenes. There has also been some research concerning high-level vision. [10] implement a global blackboard memory in a scene interpretation system, generating scene descriptions by sharing the blackboard at several abstract levels of visual interpretation. [5] uses geometric models to identify aircraft objects in aerial images of an airport scene. [2]

There has been some interest in the use of language to describe events and spatial relationships. [4] develops an event calculus which uses some low-level visual primitives to guide the interpretation of events in a robot assembly environment. [3] describes the use of spatial prepositions for generating descriptions to scenes from the viewpoint of a scene observer. [11] analyzes locative prepositions and points out that the use of such locatives establishes "ideal" relationships which must be made to fit to each particular instance of its usage. She has also pointed out that there is an implied "geometric conceptualizaton" when locatives are interpreted. [16] develops a cognitive grammar which helps to formalize the use of spatial and perceptual relationships through the use of referents and trajectors as keys which relate a linguistic grammar to the conceptualization of the objects which are spatially related. [23] explore verb-driven event processing in the observation of traffic scenes for the generation of natural language descriptions. [29] has contributed to the research with explorations of the relationship between language and spatial relations. [24] has implemented a system using visual predicates for early language development. While research has been accomplished toward understanding verbal scene description, there has not been enough work on describing the visuo-linguistic interface in terms of how vision and language influence and constrain each other to determine visual and linguistic conceptualizations.

The perceptual activities and structures associated with visual perception are not well-defined. From an apparently small set of "hard-wired" visual percepts, people seem to eventually build a relatively large set of complex visual concepts. While language helps people communicate, it is often required to also efficiently convey a large amount of perceptual information. A complete analysis of the verbal description of visual concepts would require considering the verbal communication process from perceptually low-levels through the generation of linguistic responses. This paper will concentrate however, on outlining how linguistic concepts of tense and aspect can be generated from mostly intermediate-level visual percepts.

3. Conceptualizating the Event

To further discuss the model of a visuo-linguistic interface, it is important to define what concepts and conceptualizations are. A *concept* is an association of object, state, and event (object and state changes) representations which have perceptual, linguistic, physical, and cognitive foundations. *Conceptualization* is the process of associating those representations under a common conceptual theme as concepts. There is no default structure for concepts since they are representations of distributed knowledge sources and may be associated with several other concepts. Conceptual association is constrained by the memory and processing capability of the conceptualizing agent. In this paper, we are concerned with the ***visuo-linguistic conceptualization of events:*** the process of associating sequential visual object, state, and event changes with

[1] Space does not permit us to review the low-level vision research but Cf. [2], [6], [13].

[2] An excellent collection of papers concerning computer vision systems may be found in [10].

language and vice versa. To address this concern, the description of linguistic and visual concepts must be presented. The following sections will outline linguistic and visual concepts and discuss their properties.

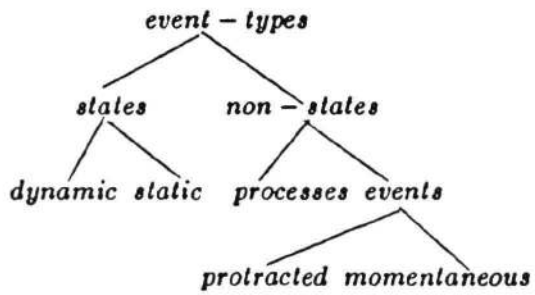
3.1 Lexical Semantics for Verbs

In this section we outline the framework that defines our domain for linguistic and lexical conceptualization. We will adopt an interval-based semantics, the *Extended Aspect Calculus* ([25]), which provides a semantics for lexical items and constrains what word meanings are possible for lexicalization in a language. The thesis of this approach is to decompose the events denoted by verbs into the subintervals that compose them (cf. [7]).

Our model is a first-order logic that employs special symbols acting as operators over the standard logical vocabulary. These are taken from three distinct semantic fields. They are: *causal*, *spatial*, and *aspectual*. The predicates associated with the causal field are: $Causer(C_1)$, $Causee(C_2)$, and $Instrument(I)$. The spatial field has two predicate types: *Locative* and *Theme*. Finally, the aspectual field has three predicates, representing three temporal intervals: t_1, beginning, t_2, middle, and t_3, end. From the interaction of these predicates all thematic types can be derived.[3]

Let us illustrate the workings of the calculus with a few examples. For each lexical item, we specify information relating to the argument structure and mappings that exist to each semantic field; we term this information the *Thematic Mapping Index (TMI)*.

Part of the semantic information specified lexically will include some classification into one of the following event-types (cf. [1], [7], [15], [26], [30]).



For example, the distinction between state, activity (or process), and accomplishment can be captured in the following way. A state can be thought of as reference to an unbounded interval, which we will simply call t_2; that is, the state spans this interval.[4] An activity or process can be thought of as referring to a designated initial point and the ensuing process; in other words, the situation spans the two intervals t_1 and t_2. Finally, an event can be viewed as referring to both an activity and a designated terminating interval; that is, the event spans all three intervals, t_1, t_2, and t_3.

We assume that part of the lexical information specified for a predicate in the dictionary is a classification into some event-type as well as the number and type of arguments it takes [18], [31]. For example, consider the verb *run* in sentence (1), and *give* in sentence (2).

(1) John ran yesterday.

(2) John gave the book to Mary.

[3] The presentation of the theory is simplified here, as we do not have the space for a complete discussion. See [25] for discussion.

[4] This is a simplication of our model, but for our purposes the difference is moot. A state is actually interpreted as a primitive homogeneous event-sequence, with downward closure. Cf. [26],

We associate with the verb *run* an aspect structure P (for process) and an argument structure of simply $run(x)$. For *give* we associate the aspect structure A (for accomplishment), and the argument structure $give(x, y, z)$. The Thematic Mapping Index for each is given below in (3) and (4).

(3)

$$run = \begin{pmatrix} C_1 \\ | \\ x \\ | \\ Th \\ \diagup \quad | \\ t_1 \quad t_2 \end{pmatrix}$$

(4)

$$give = \begin{pmatrix} C_1 \quad C_2 \\ \diagup \quad \diagup \\ x \quad y \quad z \\ | \quad | \quad | \\ L \quad Th \quad L \\ | \quad | \quad | \\ t_1 \quad t_2 \quad t_3 \end{pmatrix}$$

The sentence in (1) represents a process with no logical culmination, and the one argument is linked to the named thematic (or case) role, *Theme* [14], [27]. The entire process is associated with both the initial interval t_1 and the middle interval t_2. The argument x is linked to C_1 as well, indicating that it is an *Actor* as well as a moving object (i.e. *Theme*). This represents one TMI for an activity verb.

The structure in (2) specifies that the meaning of *give* carries with it the supposition that there is a logical culmination to the process of giving. This is captured by reference to the final subinterval, t_3. The linking between x and the L associated with t_1 is interpreted as the thematic role *Source*, while the other linked arguments, y and z are *Theme* (the book) and *Goal*, respectively. Furthermore, x is specified as a *Causer* and the object which is marked *Theme* is also an affected object (i.e. *Patient*). This will be one of the TMIs for an accomplishment.

Finally let us consider how this lexical information is actually used when we form sentences in the language. In particular, let us examine the distinction between the *simple past* forms of a sentence (5) and the *progressive* forms in (6).

(5) a. The plane landed.
b. The plane descended.

(6) a. The plane is landing.
b. The plane is descending.

Notice that (6b) entails (5b) but it is not true that (6a) entails (5a). That is, although we can say that the plane has descended if we say that it is currently descending, it is not the case that the plane has completed its landing if we say that it is landing. If we classify *descend* as an activity and *land* as an accomplishment, however, we are able to capture this distinction in entailments.

Let us say that the progressive acts as an operator over an event sequence, and picks out the middle interval t_2 as the one being referred to.

This means that the subevent being referred to by use of the progressive is inside the event t_2, and does not entail the completion of a landing, since there is no culminating event associated with the progressive. Given this analysis for the progressive, we now can explain why process verbs allow the inference *if x is V-ing, then x has V-ed.*

3.2 Visual Event Concepts

Visual event concepts (or visual events) are associations of percepts which are "hard-wired" low-level visual primitives (motion, location, intensity, size, color, etc.) and spatio-temporal relations defined by those primitives (e.g. under, near, between, etc.). A majority of percepts represent object states and relations, while only a few percepts represent events (e.g. motion). Our definitions for motion are found in Figure 1.

```
[motion
 (object (motion-right motion-left motion-forward
                       motion-backward motion-up motion-down 0))]

[motion-right
 (object (location 0) (location (x (increase 1))))]
                        ⋮
[motion-down
 (object (location 0) (location (y (decrease 1))))]
```

[Figure 1]

We assume a viewer-oriented coordinate system with the origin at the center of the field of view. The z-axis is the line of sight of the observer (+z) through the origin. The x-axis corresponds to the observers' right (+x) and left while the y-axis is up (+y) and down. The numbers are visual sampling indices which suggest the expected sequence of location states which determine motion. We present these definitions to illustrate that though motion and location can be decomposed into more primitive elements (coordinates), we will define location and motion (change of location) as our most primitive state and event, respectively. Our theory is concerned with percepts which are sufficient for verbal descriptions using tense and aspect, therefore percepts which are determined by low-level location and motion are considered intermediate-level visual percepts.

Since events are more salient than states, percepts which denote events have greater control of an observer's visual focus of attention than percepts which simply denote states. Visual events which include such percept changes can influence (support, interrupt, suspend, or terminate) the observer's attention. Furthermore, short-term memory constraints force the observer to attend to perceptual changes during an observation. We define object, state, and event changes as simple events so that a visual event may be defined as a sequence of one or more simple events. This sequence may be a sub-sequence (sub-event) of any number of other distinct visual events. For the remainder of this paper, visual events will be referred to simply as events and simple events which bound visual events will be termed initial and final events (a simple event is the initial and final event of itself).

Events are largely developed through observation which is the concurrent processing of identifying objects and their behavior, predicting and matching event-schemata, and evoking linguistic, cognitive, and physical descriptive procedures. These procedures are evoked at some level of abstraction which is appropriate for the description, which by default is the highest level. Description complexity may vary from simple perceptual recognition to combinations of linguistic, cognitive, and physical procedures. Generally, events are percept changes defined in terms of object combinations and the polyadicity (number of object arguments) of the percepts. We have identified percepts which require one, two, or three objects similar to those in [19]. Each visual sampling interval (scene) is represented by the simple event which is the set of monadic, dyadic, and triadic percepts using each object, object pair, and object triple as arguments, respectively.

The purpose of observation is to describe known event-schemata and to define new event-schemata. Generally, new schemata are constructed from a visual activity history by: retaining the history as it was observed; defining sub-events from changing "hard-wired" percepts; defining sub-events from any changing percepts or simple events; or by matching predicted simple events from known event-schemata. New event-schemata are named through interaction with a critic or by concatenating the names of previously defined events and recognized percepts. The probability of event recognition is measured by the degree to which event-schemata are matched.

4. A Visuo-linguistic Interface Model

At each scene, matched percepts, simple events, sub-events, and event-schemata will determine the generation of verbal descriptions where objects assume thematic roles. The description process is guided, in part, by recognizing whether the objects, the initial event, and the final event can be identified, partially identified, or unidentified. This will in turn determine whether the event is a definite, probable, or possible past, present, or future process, achievement, or accomplishment. The *past* is considered when all events have occurred prior to the present scene; the *present* when all simple events occur within the present scene; and the *future* is used when all events will occur after the present. The following is our algorithm for the visuo-linguistic interface:

1. **OBSERVE** $scene_n$.

When an observation begins, the observer creates a visual history in intermediate-term blackboard memory. At each scene the observer confirms the recognition of objects and the spatio-temporal relations between objects.

2. **PREDICT event-schemata.**

Predictions of long-term memory event-schemata are goal-based when selected by the observer through non-visual (e.g. verbal) input, object-based when selected by visibly identifying objects which are event-schemata agents, and event-based when selected by identifying spatio-temporal relation sequences of visibly unidentified objects and plausibly inferring event-schemata agents.

3. **DETERMINE changed percepts.**

The observer's attention is driven by percept changes during each scene. The degree of attention is roughly proportional to the number of changed percepts: the larger the number of changes, the greater the need for attention. Goal-based prediction will evoke expectation-driven attention while object-based and event-based predictions will evoke data-driven attention. Putting these together, we can define the **total attention** to be the cooperative and/or competitive interaction between the data-driven and expectation-driven mechanisms.

4. **MATCH observed sub-events with predicted event-schemata.**

Sub-events are identified by focusing attention on default "hard-wired" and/or learned percepts. Identified sub-events are matched with sub-events of predicted event-schemata.

5. **CLASSIFY predicted event-schemata using matched sub-events.**

Matched sub-events are compared with the structure of predicted event-schemata by verifying object agents, and determining the state (past, present, or future) of predicted event-schemata and their matched

sub-events. Predicted event-schemata are subsequently classified as processes, achievements, or accomplishments .

6. PRIORITIZE classified predicted event-schemata.

Predicted event-schemata are ordered based on the percept salience of the sub-events by which they were classified. For example, a nearby object quickly moving toward the observer may be more salient than a distant object moving slowly away from the observer.

7. DESCRIBE successfully predicted event-schemata.

Verbal descriptions are generated based on the salience, state, and classification of the predicted event-schemata. This will include tense, aspectual, and causal references. The fine-grained temporal granularity of visual perceptions will be mapped into medium-grained perceptual changes which are mapped into coarse-grained linguistic descriptions.

8. COMMENT and generate QUERIES about unsuccessful predictions.

Verbal comments are generated for predicted event-schemata whose descriptions suggest improbable occurrence and minimal salience. The observer will direct questions to an interactive critic in an attempt to relate successful predictions to unsuccessful predictions ([6], [21]).

9. REFINE, UPDATE, and CREATE event-schemata.

Through dialogue the observer will attempt to assign credit to percepts and sub-events in an effort to create new event-schemata and revise known event-schemata.

10. REPEAT UNTIL *scene$_f$*.

The process continues until the observation is terminated by a minimal amount of salience in the scene for an extended period of time, or through the volition of the observer.

It should be pointed out that event-schemata predictions are made in order to reduce the search problem of a large number of event-schemata with a very large number of percepts. Goal-based predictions are specific and require the less of the observer's attention resources than object-based predictions while event-based predictions are general and require more of the observer's attention than object-based predictions. This attention disparity exists at the beginning of the observation, but it is expected that by the end of the observation a small number of event-schemata will have actually been described.

5. Examples

To illustrate our theory of the integration of language and perception, consider that an observer and a critic witness air show events at an airport. There are two objects at the show: a plane and a runway. The observer is a novice and can identify planes and runways and the critic is an aviation expert. Assume that for every scene in the observation the observer perceives the location and motion of both objects. From these "hard-wired" percepts the observer determines other percepts: on, over, above, velocity, and altitude. Let us say that the observer focuses on percept changes between each scene and represents them in a visual activity history. If an observation yields the following history of percept changes:

```
(visual-history
 (runway (location 0))
 (plane  (location 0) (motion 3)
         (velocity (zero 0) (increase 4) (constant 15) (decrease 19)
                   (constant 27))
         (altitude (constant 0) (increase 10) (constant 16) (decrease 20)
                   (constant 25))
         (runway (on 0) (over above 10) (above 13) (over above 20)
                 (on 25))))
```

and the observer can verbally describe percepts, it could describe the activity in any scene in terms of the percepts:

```
(scene-0
 (runway (location 0))
 (plane  (location 0)
         (velocity (zero 0))
         (altitude (constant 0))
         (runway (on 0))))
```

$ING(sit\ x)$
$ON(Th, L)$
$Theme \rightarrow plane/[-motion]$
$L \rightarrow runway$

"A plane is sitting on a runway."

Scene 0 suggests that a motionless plane is the direct agent of sitting on a runway location. This would be the case until scene 4:

```
(scene-4
 (runway (location 0))
 (plane  (location 0) (motion 3)
         (velocity (increase 4))
         (altitude (constant 0))
         (runway (on 0))))
```

$ING(move\ x)$
$ON(Th, L)$
$Theme \rightarrow plane/[+motion]$
$L \rightarrow runway$

"The plane is moving faster on the runway."

Scene 4 shows that the plane had been on the runway since scene 0, moved since scene 3, and increased velocity in scene 4. The observer can also generate sub-events based on any particular changing percept. For instance, the observer can define a simple sub-event by focusing on the change in velocity of the plane at scene 15 and include all percept changes which occurred between the last two successive velocity changes in scenes 4 through 15 and call it a "foo":

```
(sub-event-foo
 (runway (location 0))
 (plane  (location 0) (motion 3)
         (velocity (increase 4) (constant 15))
         (altitude (increase 10))
         (runway (over above 10) (above 13))))
```

$ING(foo\ x)$
$Theme \rightarrow plane/[+motion]$
$L \rightarrow above\ runway/[+location]$

"The plane is increasing altitude above the runway at constant speed."
or
"The plane has fooed."

The observer could continue to generate descriptions of this visual activity in such terms, but for long and complex events there could be a very large number of percepts and sub-events making verbal descriptions too detailed, awkward, lengthy, or ridiculous. For these reasons, it is sometimes desirable that sub-events have more concise and meaningful descriptions. Sub-events could be identified by an interactive critic who can recognize and label them linguistically. Consider that the following dialogue takes place after witnessing the visual activity:

```
  Critic: "The plane takes-off when it accelerates on the runway
          and then ascends."
Observer: "What is ascending?".
  Critic: "The plane ascends when it increases altitude."
```

This verbal exchange causes the observer to focus attention on "increasing altitude" at scene 10. The observer now constructs an "ascend" sub-event schema:

```
(ascend
 (runway (location 0))
 (plane  (location 0) (motion 3)
         (velocity (zero 0) (increase 4))
         (altitude (constant 0) (increase 10))
         (runway (on 0) (over above 10))))
```

From the observation and the dialogue, the role of the runway in the plane's ascending is not clear. Furthermore, the critic has not given any definite indication as to when an ascend begins and ends. If the dialogue continues:

```
Observer: "When does an ascend begin?"
  Critic: "The plane begins to ascend when it increases
          altitude."
Observer: "When does it end?"
  Critic: "When the plane stops increasing altitude."
Observer: "Does a plane need a runway to ascend?"
  Critic: "No."
Observer: "Does it need velocity to ascend?"
  Critic: "Yes."
```

and scene indices are normalized, the observer may generate a more refined schema for "ascend":

```
(ascend
 (plane (location 0) (motion 1)
        (velocity (increase 2) (constant 4))
        (altitude (increase 3) (constant 5))))
```

Careful guidance by the critic could result in other refined event-schema definitions such as take-off, descend, and landing. The observer could now describe the same visual activity at a higher level of abstraction (? indicates unobserved percept):

```
(scene-4
 (runway (location 0))
 (plane  (location 0) (motion 3)
         (velocity (increase 4))
         (altitude (constant 0))
         (runway (on 0))))


(take-off
 (runway (location 0))
 (plane  (location 0) (motion 3)
         (velocity (increase 4)) (altitude (increase ?)))
         (runway (on 0) (over above ?))))
```

"The plane is taking-off."

```
(scene-23
 (runway (location 0))
 (plane  (location 0) (motion 3)
         (velocity (decrease 19))
         (altitude (decrease 20))
         (runway (over above 20))))


(descend
 (plane (location 0) (motion 3)
        (velocity (decrease 19) (constant ?))
        (altitude (decrease 20) (constant ?))))


(land
 (runway (location 0))
 (plane  (location 0) (motion 3)
         (velocity (decrease 19)) (altitude (decrease 20)))
         (runway (over above 20) (on ?))))
```

"The plane is descending and has almost landed."

In these cases, the observer is guided to define sub-events by focusing attention on suggested percepts rather than focus attention on "hard-wired" or motion-related percepts though all percepts remain building blocks for sub-events. Partial event-schemata matches were found to be helpful in generating descriptions with the use of words such as "almost" and "partially" though the events never completely occurred. The observer may now describe new visual activity in terms of events that it can recognize.

Without the benefit of instruction, it would take our observer several observations outside the proximity of an airport to notice that planes often ascend without runways and sometimes ascend due to increased wind velocity. [5] While the plane's velocity is not essential for visually recognizing ascent or descent, such percepts can be included in event-schemata to help the observer make causal inferences in verbal descriptions.

Our examples show that our visual event definitions are hierarchical (since sub-events are constructed from events) and concurrent. We are quick to point out that without the benefit of language, event boundaries may be determined by percept salience alone, however, language can help to determine and label visual events on non-salient or non-visual bases. Thus the default temporal granularity and focus of attention during event processing can be altered by using language.

6. Summary and Future Work

Our theory relates the thematic roles of objects in events to lexical and perceptual semantics. It presents a plausible mapping from visual percepts to linguistic descriptions and the inverse transformation from linguistic descriptions to visual event-schemata. We have suggested the role that language may play in describing perceptions and provide an algorithm which describes this mapping process. We introduce goal-based, object-based, and event-based prediction and show how such predictions are integrated to focus attention on input which may be linguistic as well as perceptual.

The authors would like to point out several significant directions that our research in perceptual-linguistic interfacing and related issues can be explored. First, though we are directly concerned with vision and language in this paper, such work should lead towards investigations in perceptual modality and descriptive integration. For example, the next step in defining formalisms could be to select another perceptual modality (e.g taction) and another descriptive mechanism (e.g. motor-control) and develop formalisms which describe how an intelligent, observing entity may physically move as a result of how it is physically touched. Along with the theory outlined in this paper a more complete characterization of perceptual description may result.

Another interesting avenue to explore would be how modal and descriptive integration can be controlled. One idea is that the lexicon, percepts, and event-schemata can be nodal processors in a massively parallel fine-grained computational network similar to [12] and more sophisticated memory and inference and search reduction mechanisms such as [28] may be employed. We are exploring such implementation details and find that a "Society of Mind" [20] architecture may be most promising.

[5] This is the same problem as learning the necessary conditions for an event or concept. The more general notion of the concept will arise with the right training instances. See [21], [22].

Bibliography

[1] Bach, Emmon, "The Algebra of Events", in *Linguistics and Philosophy*, 1986.

[2] Ballard, Dana H, Brown, C., *Computer Vision*, Prentice-Hall, New Jersey, 1982.

[3] Boggess, L. C., "Computational Interpretation of English Spatial Prepositions," Report T-75, Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, February 1979.

[4] Borchardt, G. C., "Event Calculus," in *Proceedings, Ninth International Joint Conference on Artificial Intelligence*, Los Angeles, August 1985, 524-27.

[5] Brooks, R. A., "Symbolic Reasoning Among 3-D Models and 2-D Images," Report STAN-CS-81-861, Department of Computer Science, Stanford University, June 1981.

[6] Cohen, P. R., and Feigenbaum, E. A., *The Handbook of Artificial Intelligence*, 3, William Kaufmann, Los Altos, 1982.

[7] Dowty, David R., *Word Meaning and Montague Grammar*, D. Reidel, Dordrecht, Holland, 1979.

[8] Fillmore, Charles, "The Case for Case", in *Universals in Linguistic Theory*, E. Bach and R. Harms (eds.). New York, Holt, Rinehart, and Winston, 1968

[9] Gruber, Jeffrey, "Studies in Lexical Relations"unpublished PhD, MIT, 1965

[10] Hanson, A. R., and Riseman, E. M., (Eds.), *Computer Vision Systems*, Academic Press, New York, 1978.

[11] Herskovits, A., "Semantics and Pragmatics of Locative Expressions," *Cognitive Science*,9, 3, 1985, 341-78.

[12] Hillis, D., *The Connection Machine*, MIT Press, Cambridge, 1985.

[13] Horn, B.K.P., *Robot Vision*, MIT Press, Cambridge, 1986.

[14] Jackendoff, Ray, *Semanic Interpretation in Generative Grammar*, MIT Press, Cambridge, MA. 1972

[15] Kenny, Arthur, *Actions, Emotions, and Will*, Humanities Press, New York. 1963

[16] Langacker, R. A., "An Introduction to Cognitive Grammar," *Cognitive Science*, 10, 1, 1986, 1-40.

[17] Marr, D., *Vision*, W. H. Freeman, San Francisco, 1982.

[18] Miller, George, "Dictionaries of the Mind" in Proceedings of the 23rd Annual Meeting of the Association for Computational Linguistics, Chicago, 1985.

[19] Miller, G. A., and Johnson-Laird, P. N., *Language and Perception*, Belknap/Harvard University Press, Cambridge, 1976.

[20] Minsky, M. L., *The Society of Mind*, Simon and Schuster, New York, 1987.

[21] Mitchell, Tom, "Version Spaces: A Candidate Elimination Approach to Rule Learning," in *Proceedings, Fifth International Joint Conference on Artificial Intelligence*, August 1977.

[22] Michalski, R.S. "A Theory and Methodology of Inductive Learning," in Michalski et al (eds.), *Machine Learning I*, Tioga Press, 1983

[23] Neumann, B., and Novak, H.-J., "Event Models for Recognition and Natural Language Description of Events in Real-World Image Sequences," in *Proceedings, Eighth International Joint Conference on Artificial Intelligence*, Karlsruhe, W. Germany, August 1983, 724-26.

[24] Pustejovsky, James, "The Acquisition of Lexical Entries: The Perceptual Origin of Thematic Relations," to appear in *Proceedings of the 25th Meeting of the Association of Computational Linguistics*, Seattle, 1987.

[25] Pustejovsky, James, "The Extended Aspect Calculus", Submission to special issue of *Computational Linguistics*, 1987

[26] Ryle, Gilbert, *The Concept of Mind*, Barnes and Noble, London, 1949

[27] Schank, Roger, *Conceptual Information Processing*, North-Holland, Amsterdam, 1975.

[28] Stanfill, C., and Waltz, D. L., "The Memory-Based Reasoning Paradigm," Thinking Machines Corporation, Cambridge, 1987.

[29] Talmy, L., "How Language Structures Space," in *Spatial Orientation: Theory, Research, and Application*, Acredolo, L., and Pick, H., (Eds.), Plenum Press, 1983.

[30] Vendler, Zeno, *Linguistics and Philosophy*, Cornell University Press, Ithaca, 1967

[31] Wilks, Yorick "Preference Semantics," *Aritficial Intelligence*, 1975.