Naive Physics, Event Perception, Lexical Semantics and Language Acquisition

1 Introduction

In a companion paper, Siskind (1991) argues that during the early bootstrapping stages of language acquisition, when children start out without knowing either syntax or the meanings of any words, children are aided in their task by hypothesizing a set of potential meanings for each utterance heard. For example, a child hearing the utterance *John entered the room* would look out into her environment and see John standing, walking, opening the door, being outside the room, and later being inside the room, along with many other possible events occurring in the environment unrelated to John. Hypothesizing that the utterance as a whole refers to one of those events aids the learner in figuring out what the individual words mean, as well as the syntactic categories of those words and the syntactic parameters of the language being learned. But how can a child hypothesize utterance meanings from visual perception? This is the topic addressed by this paper.

Since we want to understand how a child's perception of the world can aid the language acquisition task, we must look for evidence of what knowledge pre-linguistic children already possess prior to linguistic activity.[1] Spelke (1988) discusses habituation/dishabituation experiments which attempt to elucidate such knowledge. These experiments provide evidence that pre-linguistic children possess at least the following kinds of knowledge:

substantiality: the knowledge that objects take up space and cannot pass through one another,

continuity: the knowledge that an object appearing at point A and then at point B must have moved along a continuous path between those two points,

gravity: the knowledge that unsupported objects fall and

ground plane: the knowledge that the ground offers universal support for objects.

We refer to these collectively as pre-linguistic principles.

We are currently writing a program called Abigail, which attempts to incorporate such pre-linguistic knowledge into a simulated language learner to test the hypothesis that such knowledge can aid the language acquisition task. Abigail watches a computer animation constructed from line segments and circles. Along with that animation, Abigail receives a narration text describing the events occurring in the movie. The experimental paradigm of having a learner acquire new word meanings by watching a narrated movie has been explored by Rice (1990). In our case however, the learner is a machine rather than a child. Using techniques which incorporate the aforementioned pre-linguistic principles, Abigail analyzes the animation frame by frame and produces a semantic representation of the events occurring in that animation. The events of this semantic representation constitute the meanings hypothesized for utterances appearing in the narrative text. Siskind (1991) presents a learning algorithm which can utilize such a semantic representation to learn the syntactic categories and meanings of words. This paper focuses on how to produce this semantic representation from visual input using models of children's pre-linguistic knowledge.

Abigail lives in a microworld of animated movies. These movies contain objects which participate in events. The ontology of this microworld differs somewhat from that of our world. More importantly, however, the ontology of Abigail's world is similar enough to our world to model the pre-linguistic principles of substantiality, continuity, gravity and ground plane. A frame from one of Abigail's movies is shown in Figure 1. In this movie, the man walks to the table, picks up the ball, walks back and forth with it before putting it back on the table. Later, the woman repeats the same actions, and finally the man goes, picks up the ball and gives it to the woman who then puts it back on the table. Abigail's computational mechanisms are not specific to the particular objects and event in this movie. Unlike the system discussed by Badler (1975), Abigail does not possess any prior object or event models. Furthermore, the animation is generated by a program distinct from Abigail. Abigail has no access to the internal data structures of this animation program. Abigail observes only the positions, sizes, shapes and orientations of the line segments and circles comprising each animation frame. From this information, Abigail utilizes a theory based on the pre-linguistic principles of substantiality, continuity, gravity and ground plane to construct a semantic representation of the objects and events in the animation. Without any modification, Abigail can watch a different animation containing different objects participating in different events and still be able perform a semantic analysis to yield an appropriate representation of the objects and events in this new movie.

[1] This paper remains agnostic as to whether such pre-linguistic knowledge is innate or acquired during the early months of life.

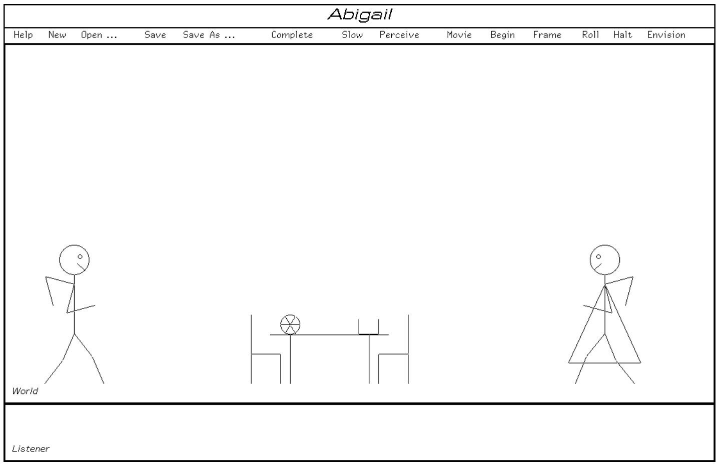

Figure 1: A frame from one of Abigail's movies.

2 The Theory

As mentioned previously, Abigail does not directly perceive objects such as people, tables and chairs. Instead Abigail perceives the figures, such as line segments and circles, out of which objects are constructed, and then interprets certain collections of figures as objects. In particular, Abigail understands that figures may be joined together. We denote a joint connecting figures f and g as $f \leftrightarrow g$. Such a joint can be described by three parameters: the displacement of the joint along the length of f, the displacement of the joint along the length of g and the angle formed between f and g. Any joint can be either rigid or flexible, independently along each of these three dimensions. A rigid joint parameter has some fixed value while a flexible joint parameter leaves its value unspecified. For technical reasons, we require that at least one of the displacement parameters of every joint be rigid. Any set of figures which are connected by joints will be interpreted as an object. Abigail does not directly perceive joints between figures. She infers those joints which are necessary to explain the unfolding animation according to the pre-linguistic principles. Furthermore, the set of joints and their parameter values need not be invariant for the duration of the movie. During the course of the movie, joints may change from rigid to flexible, or vice versa, and may even appear or disappear completely. This allows new objects to be built by combining old objects, old objects to be broken into parts and objects to be broken and then fixed again. Abigail, must continually maintain and update a joint model of the world to understand such construction and destruction events.

Abigail's microworld is nominally a two dimensional world. The figures that she perceives directly do not contain any depth information. Such a two dimensional world is not capable of supporting an interesting model of substantiality. The motion of objects in a two dimensional world which obeys substantiality is highly constrained. Nonetheless, when humans view the animation based on Figure 1 where the man walks from one side of the table to the other, they are not disturbed by the fact that in doing so, the man's figures overlap the table's figures. They never entertain the possibility that the man is walking through the table. Instead they assume that the man is walking either behind or in front of the table. In a similar fashion, Abigail attempts to reconstruct such depth information to explain the image and uphold the principle of substantiality. While not perceiving depth information directly, Abigail constructs a depth model which assigns certain figures constituting the image to the same layer and others to different layers. This model comprises a set of assertions of the form $\mathit{layer}(f) = \mathit{layer}(g)$, when figures f and g are known to be on the same layer, and $\mathit{layer}(f) \neq \mathit{layer}(g)$ when they are known not to be on the same layer. Only figures on the same layer must obey substantiality.

The layer model constitutes a partial third dimension. Abigail requires that at all times the layer model be a complete and consistent equivalence relation though not necessary total. Thus from $\mathit{layer}(f) = \mathit{layer}(g)$ and $\mathit{layer}(g) = \mathit{layer}(h)$ Abigail will infer $\mathit{layer}(f) = \mathit{layer}(h)$. Likewise, from $\mathit{layer}(f) = \mathit{layer}(g)$ and $\mathit{layer}(g) \neq \mathit{layer}(h)$ Abigail will infer $\mathit{layer}(f) \neq \mathit{layer}(h)$. However, for some pairs of layers, Abigail may not know whether or not they are on the same layer. Note that these layers are not ordered and in particular there is no notion of adjacent layers. Additionally, the assignment of figures to layers may change during the course of the movie. Thus Abigail must continually update the layer model both to maintain its internal consistency as well as to uphold substantiality judgments in the changing world.

The layer model consists of a list $(a_1, \ldots, a_n)$ of layer assertions. New assertions are always added to the front of this list. Whenever new assertions are added, we check the consistency of successively longer initial prefixes of the model. If the prefix $(a_1, \ldots, a_{i-1})$ is consistent but the prefix $(a_1, \ldots, a_i)$ is not, then the assertion a_i is removed from the model. This is repeated until the entire model is consistent.

How does Abigail apply the pre-linguistic principles to update both the layer and joint model? At every frame, Abigail looks for six types of evidence between every pair of figures f and g.

1. Evidence that the assertion $\mathit{layer}(f) = \mathit{layer}(g)$ should be added to the model.

2. Evidence that the assertion $\mathit{layer}(f) \neq \mathit{layer}(g)$ should be added to the model.

3. Evidence that some parameter of the joint $f \leftrightarrow g$ should be demoted from rigid to flexible.

4. Evidence that an existing joint $f \leftrightarrow g$ should be removed from the model.

5. Evidence that some parameter of the joint $f \leftrightarrow g$ should be promoted from flexible to rigid.

6. Evidence that a new joint $f \leftrightarrow g$ should be added to the model.

Two forms of evidence can be used to infer case 1: support and collision. Whenever two figures touch and one would fall without being supported by the other, Abigail can infer that they are on the same layer. Likewise, if one figure moves toward another figure, touches it, and moves away from it according to the laws of physics, the apparent collision gives evidence that the two figures are on the same layer. Collision detection is not currently implemented in Abigail. In a similar fashion, there are two forms of evidence for case 2: overlap and exiting an apparent container. A direct observation that two figures overlap give clear evidence that they are on different layers. Furthermore, if one figure is initially surrounded by another figure and then moves so that it is no longer surrounded by that figure, the principles of continuity and substantiality imply that those two figures must be on different layers. Currently, only direct observation is implemented in Abigail. For case 3, an observation that the value of some rigid parameter of a joint has changed is evidence for demoting that parameter. An observation that two figures no longer intersect is evidence for case 4. Abigail currently does not implement any evidence for case 5. For case 6, Abigail infers a new joint whenever two figures touch and the two figures would cease to touch under the effect of gravity if they were not connected by a joint. In general, whenever Abigail hypothesizes new joints and same layer assertions to account for the stability of an object in the image, she attempts to hypothesize a minimal set of new joints and same layer assertions with same layer assertions taking priority over new joints when both offer the same explanatory power.

Central to the above process is a mechanism for determining support relationships between objects. Abigail uses a simulator for this purpose. This simulator takes the figures appearing in the current frame, along with a set of joints and layer assertions, and predicts how the image will change under the effect of gravity. This simulator is essentially a quantitative kinematic simulator that incorporates the pre-linguistic principles of substantiality, continuity, gravity and ground plane. It lacks any notion of dynamics, such as momentum, kinetic energy and friction. Nonetheless, it is adequate for determining the support relationships between objects, the same layer relationships between figures and the necessity of joints between figures.

Abigail continually performs such simulations every frame, hypothesizing what would happen in the world under different sets of joint and layer assertion assumptions. This has fairly strong psychological implications. For Abigail to be a plausible reflection of human perception, humans must be shown to be capable of performing such simulations and must also be shown to be performing them fairly regularly, albeit subconsciously. Freyd, Pantzer and Cheng (1988) gives evidence that humans perceive objects to displace slightly downward, as if they were falling, when support is removed from them.

Once Abigail has constructed the joint and layer model for each frame, and has collected connected figures into objects, she computes the following relations between those objects and the regions of space that they occupy:

$[i,j]$***exists***(α)**:** Object α exists continually for frames i through j.

$[i,j]$***contacts***(α,β)**:** Object α touches and is on the same layer as object β continually for frames i through j.

$[i,j]$***joined***(α,β)**:** For frames i through j, objects α and β are joined together by at least one joint connecting a figure from α to a figure from β.

$[i,j]$***supports***(α,β)**:** For frames i through j, object β falls if the image is simulated without object α but object β does not fall if the image is simulated with object α.

$[i,j]$***supported***(α)**:** For frames i through j, object α does not fall when the image is simulated.

$[i,j]$***moving***(α)**:** For every frame between i and j, the position, size or orientation of some figure in object α has changed from the previous frame.

$[i,j]$***moving-root***(α)**:** For every frame between i and j, the position, size or orientation of some figure in the root of object α has changed from the previous frame. The root of an object is defined to be the subset of its figures which has the greatest mass and which is connected by joints which have not changed parameters since the previous frame.

$[i,j]$***translating***(α,p)**:** Indicates that the center of mass of the root of object α is changing position for every frame between i and j. The path p is a trace of the movement of that center of mass.

$[i,j]$***rotating-clockwise***(α)**:** The root of object α is rotating clockwise for every frame between i and j.

$[i,j]$***rotating-counterclockwise***(α)**:** The root of object α is rotating counterclockwise for every frame between i and j.

$[i,j]$***rotating***(α)**:** For frames i through j, the root of object α is rotating either clockwise or counterclockwise.

$[i,j]$***place***(α,p)**:** Object α occupies the region of space indicated by p for frames i through j.

at(p,q)**:** Points p and q are approximately coincident modulo a tolerance.

in(p,q)**:** Region p is a subregion of region q.

to(p,q)**:** The ending point of path p is approximately coincident with point q modulo a tolerance.

from(p,q)**:** The starting point of path p is approximately coincident with point q modulo a tolerance.

towards(p,q)**:** Every point along path p is closer to point q than the previous point along that path.

away-from(p,q)**:** Every point along path p is further away from point q than the previous point along that path.

up(p)**:** The y-coordinate of every point along path p is greater than the y-coordinate of the previous point along that path.

down(p)**:** The y-coordinate of every point along path p is less than the y-coordinate of the previous point along that path.

3 An Example

The above relations are the primitives out of which semantic representations of events are constructed. Consider an event such as *John kicked the ball in the room.* This event could be represented as follows using the above primitives:

$$\begin{gathered}[t_1,t_2]\mathit{translating}(\mathit{foot}(\mathbf{John}),p_1)\wedge[t_1,t_2]\mathit{place}(\mathbf{ball},p_2)\wedge\mathit{towards}(p_1,\mathit{center\text{-}of\text{-}mass}(p_2))\wedge\\ [t_2,t_2]\mathit{contacts}(\mathit{foot}(\mathbf{John}),\mathbf{ball})\wedge[t_2,t_3]\mathit{translating}(\mathbf{ball},p_3)\wedge\\ [t_3,t_4]\mathit{place}(\mathbf{ball},p_4)\wedge[t_1,t_4]\mathit{place}(\mathbf{room},p_5)\wedge\mathit{in}(p_4,p_5)\end{gathered}$$

Each of the relations in the above expression can be derived from an animation of this event using the techniques described in this paper. A future paper will discuss how these relations are aggregated together to form the composite event description and how such an event description can be used by a language learner to learn the meanings of words in an utterance describing that event.

References

[1] Norman I. Badler. Temporal scene analysis: Conceptual descriptions of object movements. Technical Report 80, University of Toronto Department of Computer Science, February 1975.

[2] Jennifer J. Freyd, Teresa M. Pantzer, and Jeannette L. Cheng. Representing statics as forces in equilibrium. *Journal of Experimental Psychology, General*, 117(4):395–407, December 1988.

[3] Mabel Rice. Preschoolers' QUIL: Quick incidental learning of words. In G. Conti-Ramsden and C. E. Snow, editors, *Childrens Language (Vol. 7)*, chapter 8, pages 171–195. Lawrence Erlbaum Associates, Hillsdale, NJ, 1990.

[4] Jeffrey Mark Siskind. Dispelling myths about language bootstrapping. In *The AAAI Spring Symposium Workshop on Machine Learning of Natural Language and Ontology*, pages 157–164, March 1991.

[5] Elizabeth S. Spelke. The origins of physical knowledge. In L. Weiskrantz, editor, *Thought Without Language*, chapter 7, pages 168–184. Oxford University Press, New York, NY, 1988.