Maps for Verbs: The Relation Between Interaction Dynamics and Verb Use

Paul R. Cohen and **Clayton T. Morrison**
Center for Research on Unexpected Events (CRUE)
USC Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, California 90292
{cohen,clayton}@isi.edu

Erin Cannon
Department of Psychology
University of Massachusetts
Amherst, Massachusetts
ecannon@psych.umass.edu

Abstract

We report a study of word meaning that tests whether dynamical aspects of movies predict word use. The movies were based on a novel representation of verb semantics called *maps for verbs*. We asked preschool-school-age children to describe the movies, and demonstrated that their distributions of words could be predicted by the dynamical aspects of the movies. These results lend support to the empiricist position that word meanings are learned associatively.

1 Introduction

Previous work has shown that robots can learn the meanings of words by associating aspects of the perceptual array with utterances (e.g., [Steels, 1999; Oates, 2002; Siskind, 2001; Vogt, 2002; Cohen and Oates, 1998]). In this paper we test the conjecture that the dynamics of the perceptual array, specifically, the manner in which objects move and interact, explains the production of particular verbs. The work is part of a larger project to model human language development on robot platforms, and is based on the *maps for verbs* framework described in Section 3. A persistent question in the project is whether the perceptual array contains enough information to provide semantics for words, or, in a slightly different formulation, what fraction of the variability of word use is explained by information in the perceptual array? A concrete version of this question is posed here: What fraction of the variability of word use is explained by the dynamical aspects of interactions between two objects?

We developed 18 movies of interactions between objects. Each movie was generated by a program written in *breve* [Klein, 2002], an animation tool with good physics. A vector of parameters of each program characterizes the dynamics of the corresponding movie. We showed the movies to preschool children and asked them to describe the action in the movies. After removing non-content words, we characterized each movie by a distribution of words. We ranked all pairs of movies in two ways: by the similarities of their word distributions and by the similarities of their vectors of program parameters. Then we compared the rankings. The results are highly significant: there is strong dependence between the parameters of the programs that generated the movies and the distributions of words that children use to describe the movies.

2 Related Work

Interest in the perception of the dynamics of whole-body interactions is not new. Heider and Simmel [1944] report a study in which adults were shown a film of animated shapes interacting with each other in and around a box. After watching the film, participants were asked to describe what happened in the film. Heider and Simmel found a strong tendency to attribute a rich set of intentions to the moving objects and a story-line describing the interactions, even though the only information in the stimuli was the shape and size of the objects and their motion dynamics.

More recently, several psychologists and computer scientists have explored dynamic maps as representations of activities, often for machine recognition and modeling of human gesture and bodily movement (e.g., [Thelen and Smith, 1994; Rosenstein, 1998; Intille and Bobick, 1999; Bobick and Davis, 2001].

Blythe, Todd & Miller [1999] and Todd & Barrett [2000] present a preliminary study of adult and child perception of intention based on the dynamics of motion between two simple interacting bodies. Their work also uses the tools of dynamic map representation to characterize interactions. Their results suggest that such dynamics are implicated in people's categorization of interactions, although their focus was on intention, rather than more primitive verb classes.

For a comprehensive review of the psychological literature on dynamics and word meanings see [Cannon and Cohen, 2005].

3 Maps for Verbs

In the *maps for verbs* representation of verb meanings, the denotations of verbs dealing with interactions between two bodies, such as push, hit, chase, and so on, are represented as pathways through a metric space, or map, the axes of which are perceived distance, velocity, and energy transfer [Cohen, 1998]. Verbs with similar meanings have similar pathways. A scene, such as one object chasing another, is thought to be perceived as a pathway through the map. To learn verb

meanings, one simply associates verbs that describe scenes with the corresponding pathways.

Although maps are compact and objective representations of some verb meanings, we do not know whether they have psychological reality — whether humans use maps to assign meanings to verbs. Even if they do, the original maps for verbs representation might have the wrong axes, or the axes might be correct but verbs might not be correlated with the particular features of pathways, as we thought. The experiment in this paper does not test whether humans have maps in their heads. Instead it asks, "If one creates movies which are different according to the maps for verbs framework, will human subjects use different distributions of words to describe them?"

In the maps for verbs framework, the dynamics of interaction are split into before, during (contact), and after phases. Figure 1 depicts these phases with illustrative trajectories in eac. The axes of the maps are the same in the *before* and *after* phases: they are relative velocity and distance between the two bodies. Relative velocity is the difference between the velocity of one body, A, and another, B: $Velocity(A) - Velocity(B)$. Many verbs (e.g., transitive verbs) predicate one body as the "actor" and the other as the "target" (or "subject" or "recipient") of the action. For example, in a push interaction, the actor does the pushing, and the target is the body being pushed. By convention, the actor is designated as A and the target is B. Thus, when relative velocity is positive, the actor's velocity is greater than that of the target; and when relative velocity is negative, the target's velocity is greater than that of the actor. Distance, in turn, is simple Euclidean distance between the bodies.

The vertical dimension of the map in the *during* phase is perceived energy transfer (from the actor to the target). If energy transfer is positive, then the actor is imparting to the target more energy than the target originally had; if energy transfer is negative, then the situation is reverse and the target is imparting more energy to the actor. Since energy transfer is not directly perceivable, we approximate it by calculating the acceleration of the actor in the direction of the target while the actor and target are in contact.

The labeled trajectories in Figure 1 characterize the component phases of seven interaction types, as described by the verbs push, shove, hit, harass, bounce, counter-shove and chase.

For example, $\langle \mathbf{b}, \mathbf{b}, \mathbf{b} \rangle$ describes a *shove*. The actor approaches the target at a greater velocity than the target, closing the distance between the two bodies. As it nears the target, the actor slows, decreasing its velocity to match that of the target. Trajectory **b** of the before phase in Figure 1 illustrates these dynamics, showing the decrease in relative velocity, along with decrease in distance. At contact, the relative velocity is near or equal to zero. During the contact phase, the actor rapidly imparts more energy to the target in a short amount of time, as illustrated by **b** of the during/contact phase. And after breaking-off contact with the target, the agent rapidly decreases its velocity while the target moves at a greater velocity due to the energy imparted it.

In Figure 2(b), below, we provide a plot of the dynamics of a simulated shove action. The map in the figure plots the dynamics for a portion of the time between contact phases. The trajectory begins with very low relative velocity, as would be expected just after completing the contact phase of a shove (after phase b in Figure 1), and ends with a high relative velocity that is ramping down (before phase b in Figure 1) just before a new shove occurs.

With this three-phase representation scheme, we define six more interaction types corresponding to common English verbs:

- *Push* $\langle \mathbf{b}, \mathbf{a}, \mathbf{a} \rangle$ – begins like shove, but at contact relative velocity is near or equal to zero and the actor smoothly imparts more energy to the target; after breaking contact, the agent gradually decreases its velocity.
- *Hit* $\langle \mathbf{c}/\mathbf{d}, \mathbf{c}, \mathbf{c} \rangle$ – may begin with the actor already at high velocity relative to the target or increasing in relative velocity, and thus is characterized by **c** or **d** in the before phase.
- *Harass* $\langle \mathbf{c}/\mathbf{d}, \mathbf{c}, \mathbf{d} \rangle$ – is similar to a hit, except the after-phase involves the actor quickly recovering its speed and moving back toward the target, not allowing the distance between the two to get very large (the **d** in the after phase). Harass highlights that interactions may be cyclic: the after phase of one epoch blends into the before phase of the next.
- *Bounce* $\langle \mathbf{c}/\mathbf{d}, \mathbf{d}, \mathbf{e} \rangle$ – along with counter-shove, bounce involves the target making a more reactive response to the actor's actions. Bounce begins like a hit or harass, but at contact, the target transfers a large amount of energy back to the actor.
- *Counter-shove* $\langle \mathbf{b}/\mathbf{c}/\mathbf{d}, \mathbf{e}, \mathbf{e} \rangle$ – is a version of a shove where the target imparts energy to the actor.
- *Chase* $\langle \mathbf{a}, -, - \rangle$ – involves the actor moving toward the target, closing the distance between the two, but never quite making contact, so the during and after phases are not relevant. This is depicted as the circular trajectory **a** in the before phase.

4 Experiment

4.1 Stimuli

We used *breve 1.4*, an environment for developing realistic multi-body simulations in a three dimensional world with physics [Klein, 2002], to implement a model of the seven interaction classes described in the previous section. The models were rendered as two generic objects (a blue ball for the actor and a red ball for the target) moving on a white background — see Figure 2(a). The models allowed us to generate multiple instances of each interaction type.

We generated a set of movies based on each of the *breve* interaction models. For several of the interaction classes we also varied the behavior of the target object, as follows: the target object, (a) did not move except when contacted ("stationary"), (b) moved independently in a random walk ("wander"), or (c) moved according to billiard ball ballistic physics, based on the force of the collision ("coast"). We generated a total of 17 unique movies. For the bounce and counter-shove interaction types, we only implemented "stationary"

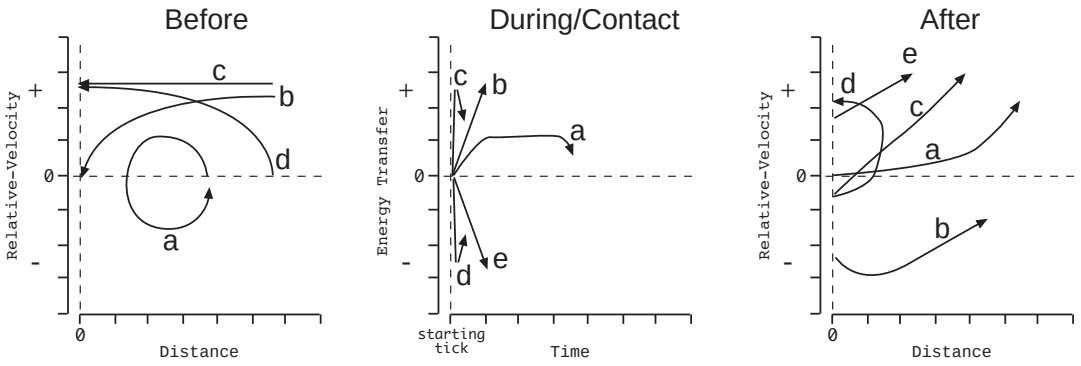

Figure 1: Maps-for-verbs model of the three phases of interaction.

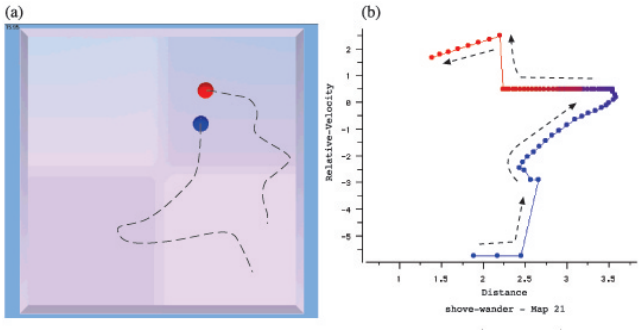

Figure 2: (a) Example of maps-for-verbs simulation running the shove-wander action, as rendered in *breve 1.4*. (Note: dashed lines represent motions of colored patches for demonstration purposes; only the moving color-patches themselves were displayed in the stimuli movies.); (b) Dynamic map plot of shove-wander action before contact, corresponding to the picture in (a) (x-axis = distance between agents, y-axis = relative velocity).

and "wander" target behavior, as "coast" would obliterate the effect of the target transferring energy back to the actor. Also, there was only one version of "chase," as the target must always be moving away from the actor. Chase was also unique because it was the only instance in which the two balls never contacted each other.

The 17 movies were recorded and presented on a G3 iMac with 14 inch screen. The children's responses were recorded and later transcribed.

Participants

Sixteen children participated in this study, ranging in age from 26-60 months old (average age = 50 months). Participants were recruited and tested at a local daycare in Amherst, MA.

Procedure

For each child, a total of 18 movies was presented, each movie instance appearing once — with the exception of chase, which the child watched twice. An experimenter told each child that she would be watching movies on the computer screen with two balls, one blue and one red, and that the task was to tell a story about what the balls were doing.

4.2 Analysis

The children used remarkably small vocabularies and very terse sentences to describe the movies; the following transcript is typical:

E: Okay, last one. Can you tell me a good story about this one?
S: Even gooder than all of the other ones?
E: Make it the best story!
S: It's going umm gooder and it's playing but the red is letting the blue push him. And the red is letting the blue one push
E: How come he's letting the blue push?
S: Because he wanted to.
E: Why does he want to?
S: Because he likes to play like that.

All the content words for each trial were extracted and "canonicalized," converting verbs in different tenses or forms (e.g., ending in -ed, -ing, etc.) to a single form. Also, negation phrases, such as "it's not zooming" or "red didn't move," were also transformed into a single token, e.g., not-zooming and not-moving. The total number of unique, canonicalized content words uttered by all the subjects in response to all the movies was 104, of which the following 30 words were uttered more than three times (words are listed with their frequencies): PUSHING, 85; MOVING, 57; BONKING, 54; AWAY, 46; TRYING, 38; PLAYING, 28; FAST, 27; RUNNING, 27; AROUND, 25; UP, 25; GETTING, 21; CHASING, 19; FRIENDS, 18; BUMPING, 17; SLOW, 16; HITTING, 16; DOWN, 16; CIRCLE, 11; CATCHING, 10; STANDING, 7; TAG, 7; ZOOMING, 6; STOPPING, 6; COMING, 4; FLYING, 4; KNOCKING, 4; FOLLOWING, 4; BOUNCING, 4; ABOUT, 4; TOGETHER, 4.

Note that some of these words are not verbs; for instance, "friends" and "away."

Each movie is characterized by a vector of relative frequencies of these 30 words. For example, here is the vector for the movie SHOVE-STATIONARY:

PUSHING, 0.192; MOVING, 0.115; BONKING, 0.0; AWAY, 0.115; TRYING, 0.038; PLAYING, 0.0; FAST, 0.115; RUNNING, 0.077; AROUND, 0.0; UP, 0.0; GETTING, 0.0; CHASING, 0.038;

:BEFORE-SUBJECT-ACTION	'agent-coast
:BEFORE-ACTOR-ACTION	'approach-slow-down
:BEFORE-ACTOR-ORIGINAL-SPEED	2
:BEFORE-ACTOR-DESIRED-SPEED	.5
:BEFORE-ACTOR-DISTANCE-FOR-SLOW-DOWN	1.5
:BEFORE-ACTOR-DISTANCE-FOR-STOP	0
:DURING-ACTOR-ACTION	'agent-approach-speed-up
:DURING-ACTOR-SPEED-UP-START-TIME	'self
:DURING-ACTOR-SPEED-UP-END-TIME	.4
:DURING-ACTOR-ORIGINAL-SPEED	'self
:DURING-ACTOR-DESIRED-SPEED	6
:DURING-ACTOR-SELF-WAIT	.6
:DURING-2-ACTOR-ACTION	'agent-halt
:DURING-2-ACTOR-SELF-WAIT	.5
:AFTER-SUBJECT-ACTION	'agent-coast
:AFTER-ACTOR-ACTION	'agent-run-away-speed-up
:AFTER-ACTOR-SPEED-UP-START-TIME	'self
:AFTER-ACTOR-SPEED-UP-END-TIME	1
:AFTER-ACTOR-ORIGINAL-SPEED	'self
:AFTER-ACTOR-DESIRED-SPEED	2
:AFTER-2-SUBJECT-ACTION	'agent-halt
:AFTER-2-ACTOR-SELF-WAIT	1

Table 1: A vector of parameters for the movie SHOVE-STATIC

FRIENDS, 0.077; BUMPING, 0.038; SLOW, 0.038; HITTING, 0.0; DOWN, 0.0; CIRCLE, 0.0; CATCHING, 0.038; STANDING, 0.0; TAG, 0.0; ZOOMING, 0.038; STOPPING, 0.0; COMING, 0.0; FLYING, 0.0; KNOCKING, 0.0; FOLLOWING, 0.0; BOUNCING, 0.038; ABOUT, 0.038; TOGETHER, 0.0.

That is, of the content words used to describe the movie SHOVE-STATIONARY, 19% of them were PUSHING, 11.5% were MOVING, and so on. Let V_{words}(movie) denote the vector of relative frequencies of words used to describe a movie.

The movies also can be characterized by the parameters of the *breve* programs that generated them. These parameters include the current speed and desired speed of the objects, the subroutines that implement patterns of movement, the distance between objects at which one slows down, the latency before starting the next movement, and so on. Table 1 shows the vector for the movie SHOVE-STATIC. Let V_{params}(movie) denote the vector of parameters for the program that generates the movie.

The next step of the analysis is to test whether there is an association between the vectors of parameters for movies, V_{params}(movie), and the vectors of words used to describe the movies, V_{words}(movie). A simple method is to calculate the similarities between pairs of movies i and j, rank the pairs by similarity, and see whether the ranking based on word vectors is predicted by the ranking based on parameter vectors. Let

$$Sim_{words}(i,j) = f(V_{words}(i), V_{words}(j)) \quad (1)$$
$$Sim_{params}(i,j) = g(V_{params}(i), V_{params}(j)) \quad (2)$$

where f and g are methods for comparing word-frequency vectors and parameter vectors, respectively. The question is whether, averaged over pairs of movies i and j, $Sim_{words}(i,j)$ is predicted by $Sim_{params}(i,j)$. We let f be generalized Euclidean distance between the probabilities $p(w_i)$ and $p(w_j)$ of hearing word w uttered in response to movies i and j, respectively:

$$f(i,j) = \sqrt{\sum_{w \in \{pushing, moving, ..\}} (p(w_i) - p(w_j))^2} \quad (3)$$

The function g, for comparing parameter vectors, cannot be simple Euclidean distance because parameter vectors include non-numeric values. We wrote a function that increases the similarity score when symbolic (and numeric values) are identical, decreases the score by a constant when symbolic values don't match, and decreases the score proportional to the mismatch between numeric values. We added some conditions for missing values. We are aware that this function might conceivably be "tuned" to make the parameter vectors better predict the word vectors, so we wrote it once and did not revise it. The results presented below are for the first and only evaluation of this function on these data.

As a final step, we ranked pairs of movies according to $Sim_{words}(i,j)$ and $Sim_{params}(i,j)$.

4.3 Results

Each pair of movies i, j gets two similarity scores, $Sim_{words}(i,j)$ and $Sim_{params}(i,j)$, so we can look at the simple correlation of these scores and the regression of $Sim_{params}(i,j)$ on $Sim_{words}(i,j)$. Figure 3 shows these score plotted against each other and the linear regression line that fits them best. Clearly, the similarity of two movies according to their parameters is a good predictor of the similarity of distributions of words used to describe the movies. The correlation between them is 0.949, which means the similarity of two movies according to their parameters accounts for 90% of the variance in the similarity of the word vectors for the movies. This result is highly significant ($p < .0001$).

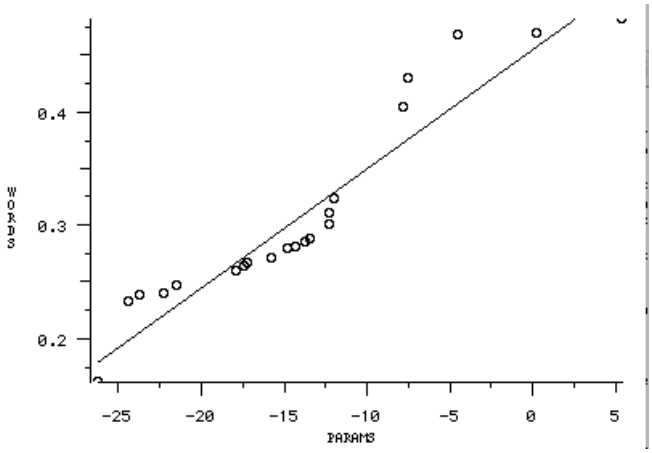

Figure 3: Regression plot of similarity scores for movie pairs according to movie parameters and movie word descriptions.

Another way to analyze the data is to compare the rankings of pairs of movies according to the two similarity scores. To illustrate, suppose we have four movies, a, b, c, d, and so six pairs of movies, ranked by function f from most to least similar: $((a,b)(a,c)(b,d)(a,d)(b,c)(c,d))$. Now suppose function g produces a different ranking: $((b,d)(a,c)(a,d)(c,d)(b,c)(a,b))$. According to f, the most similar movies, with rank 1, are (a,b) whereas the rank of these movies is 6 according to g. We can characterize the difference of two rankings by summing the differences in rank

over items; for instance, the item (a, b) contributes $6 - 1 = 5$ to the sum.

For the data in Figure 3, the summed rank difference is $\delta = 104$, but how can we tell whether this is a statistically significant number? We perform a *randomization test* as follows: There are 21 pairs of movies, so construct two vectors, v_1 and v_2, each containing the numbers $1...21$. Now shuffle v_2 thoroughly and calculate $\delta* = \sum_{i=[0,21]} abs(v_{1,i} - v_{2,i})$. Repeat 1000 times. The resulting distribution of $\delta*$ is the *empirical sampling distribution* of the summed rank difference under the *null hypothesis* that the ranks of items in the vectors are unrelated. If the actual summed rank difference, δ, has low probability according to this sampling distribution, we reject the null hypothesis with residual uncertainty (the p value) equal to the quantile of δ in the sampling distribution (see [Cohen, 1995] for details).

As it happens, the ranking of movie pairs according to the movie parameters is not independent of the ranking according to word frequencies. The summed rank difference $\delta = 104$ is just the 0.0016 quantile of the sampling distribution, so we can reject the null hypothesis and conclude with confidence that the rankings are related.

5 Discussion

These results are supported by another study, using the same movies, with adult subjects. (The final paper will present both sets of results.) They show conclusively that the dynamics of interactions between two bodies, as represented by the parameters of the programs that generate movies, predict the distributions of words that children use to describe the movies. This result is surprisingly strong when one considers how few words children actually use. Although our study involved 16 children, only 30 content words were uttered more than three times in the entire study. Nevertheless, by concentrating on the *distributions* of these words for each movie we were able to show a strong dependence on the dynamics of the movies.

Returning to the question that motivates this work, is there sufficient information in the perceptual array, particularly dynamical information, to supply semantics for some words, particularly some verbs? By demonstrating a dependency between dynamical information and word choice (especially in young children) we strengthen the empiricist case that word meanings can be learned as associations between the words and percepts. Our evidence is only suggestive, however, because we demonstrated a dependence of word use not on the child's *percepts* but, rather, on the parameters of the movies shown to the child. Only by assuming that these parameters affect how the movie is perceived can we argue for associative learning of word meanings. The assumption is very reasonable; after all, the parameters were tuned to make the movies look different and distinctive.

Still, we think it likely that associative learning of word meanings "needs help," probably from prior domain knowledge. Although researchers such as Todd and Barrett [Todd and Barrett, 2000] argue that motion is a cue to intention, they do not argue that it is the *only* cue, and one is hard-pressed to see how the intentional aspect of, say, "chasing," can be learned from nothing but the relative motions of two bodies. We think Dennett [Dennett, 1987] is probably right when he says we adopt an intentional stance even to non-intentional scenarios, and the intentional language of our subjects, directed to a couple of colored blobs moving on the screen, lends force to his argument. It seems likely to us that perceived movement is one cue to word meaning, but the intentional aspects of words are generated by the subjects themselves. We are currently designing studies to tease apart these contributors to word meanings.

Finally, we relate an anecdote about our word-learning robot: It learned that "forward" is associated with the wheels rotating in one direction and "backward" with the opposite direction of rotation, but it never learned that forward and backward are antonyms. We realized then that word meanings can be arbitrarily "deep" and that associating percepts and words produces a shallow kind of meaning. But it's a start.

References

[Blythe *et al.*, 1999] P. W. Blythe, P. M. Todd, and G. F. Miller. *How motion reveals intention: Categorizing social interactions*, pages 257–285. Oxford University Press, New York, NY, 1999.

[Bobick and Davis, 2001] A. Bobick and J. Davis. The recognition of human movement using temporal templates. *IEEE Transaction on Pattern Analysis & Machine Intelligence*, 23(3), 2001.

[Cannon and Cohen, 2005] E.N. Cannon and P.R Cohen. *Word choice as a conditional probability: Motion-based semantics*. Oxford University Press, New York, NY, 2005.

[Cohen and Oates, 1998] P. R. Cohen and T. Oates. A dynamical basis for the semantic content of verbs. In *Grounding of Word Meaning: Data & Models Workshop, AAAI-98*, pages 5–8, 1998.

[Cohen, 1995] Paul R. Cohen. *Empirical Methods for Artificial Intelligence*. The MIT Press, Cambridge, 1995.

[Cohen, 1998] P. R. Cohen. Maps for verbs. In *Proceedings of the Information and Technology Systems Conference, Fifteenth IFIP World Computer Congress*, pages 21–33, 1998.

[Dennett, 1987] D. Dennett. *The Intentional Stance)*. MIT Press, 1987.

[Heider and Simmel, 1944] F. Heider and M. Simmel. An experimental study of apparent behaviour. *American Journal of Psychology*, 57(2):243–59, 1944.

[Intille and Bobick, 1999] S. Intille and A. Bobick. A framework for recognizing multi-agent action from visual evidence. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pages 518–525, Orlando, FL, July 1999.

[Klein, 2002] J Klein. breve: a 3d simulation environment for the simulation of decentralized systems and artificial life. In *Proceedings of Artificial Life VIII, the 8th International Conference on the Simulation and Synthesis of Living Systems*, 2002. http://www.spiderland.org/breve/.

[Oates, 2002] T. Oates. Peruse: An unsupervised algorithm for finding recurring patterns in time series. In *Proceedings of the Nineteenth International Conference on Machine Learning*, 2002.

[Rosenstein, 1998] M. T. Rosenstein. Concepts from time series. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence*, pages 739–745, 1998.

[Siskind, 2001] J. Siskind. Grounding lexical semantics of verbs in visual perception using force dynamics and event logic. *Journal of AI Research*, 15:31 – 90, 2001.

[Steels, 1999] Luc Steels. *The Talking Heads Experiment: Volume I. Words and Meanings*. Laboratorium, Antwerpen, 1999. This is a museum catalog but is in preparation as a book.

[Thelen and Smith, 1994] E. Thelen and L. Smith. *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press, Cambridge, MA, 1994.

[Todd and Barrett, 2000] P.M. Todd and H. C. Barrett. Judgment of domain-specific intentionality based solely on motion cues. Paper presented at the 12th Annual Meeting of the Human Behavior and Evolution Society, 2000.

[Vogt, 2002] P. Vogt. Anchoring symbols to sensorimotor control, 2002.