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

Author(s) removed for blind review

Keywords: cognitive modeling, perception

Abstract

The *maps-for-verbs* framework predicts that our use of verbs to describe simple whole-body interactions is influenced by the characteristics of the physical dynamics in the before, during and after phases of interaction. We report two studies in which young children and adults were asked to describe movies of simple interactions governed by the dynamics proposed in the maps-for-verbs framework. The results suggest that physical dynamics do influence child and adult verb use. We discuss what we learned and our plans for future study.

Introduction

Nativists and constructivists alike share the belief that humans are born with or acquire a core set of concepts that serve as the foundation for later knowledge. Whether innate or developed, the core semantics is inevitable, a function of needing to negotiate the physical world in which distinguishing between the dynamics of physical interactions is a prerequisite to meeting goals and surviving. A number of cognitive scientists have specifically built theories of the genesis of this core semantics as deriving from interaction with the physical world (Barsalou 1999; Johnson 1987; Lakoff 1987; Mandler 1992). The work we present in this paper is part of a larger project aimed at accounting for this semantic core, with a particular focus on the prerequisites for language learning [reference removed]. In this paper we present two studies that assess whether a particular representation of verb meanings explains human verb use.

In [reference removed]'s *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, can be represented as pathways through a metric space, or map, the axes of which are perceived distance, velocity, and energy transfer. 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 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. We report the results of two preliminary studies of the use of this framework to predict the verb use of children and adults describing simple whole-body interactions.

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 colored 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 shapes and colors of the objects and their motion dynamics.

More recently, [reference removed] have explored the use of dynamic maps as a suitable representation of activities. This work developed a set of techniques, adopted from nonlinear dynamics research, to analyze and build classifiers of dynamic map patterns. This work also developed the foundation of the maps-for-verbs framework.

Bobick (Intille & Bobick 1999; Bobick & Davis 2001) has also developed the use of dynamic maps – "temporal templates" – and other techniques for the machine recognition and modeling of human gesture and bodily movement.

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.

The Maps-for-verbs framework

[reference removed] proposes maps for verbs as a possible semantics for verbs describing interactions between two objects. According to the model, the dynamics of interaction are split into before, during (contact), and after phases.

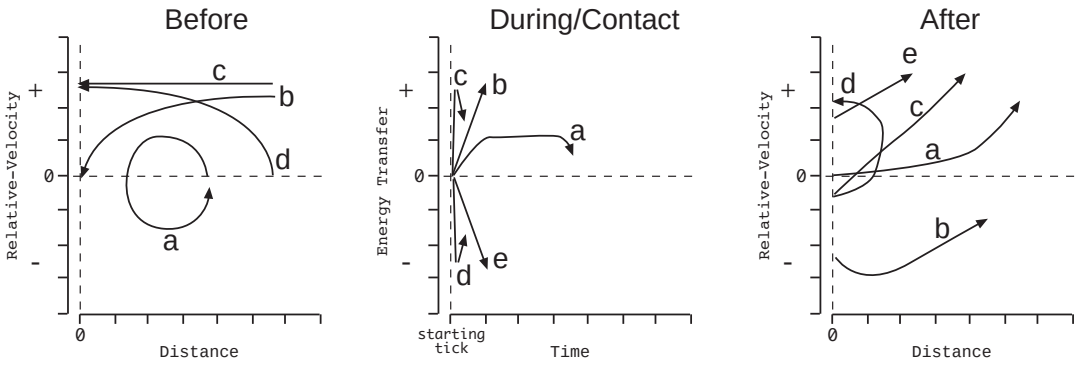

Figure 1: Maps-for-verbs model of the three phases of interaction.

These phases are characterized by relating dynamic features of interaction, such as relative velocity and distance, to define a metric space called a map (also called a phase portrait or phase diagram). A map portrays the changes in the relationships between the two bodies over time. A given interaction is then described as a directed pathway or trajectory through the map's dynamics space. Figure 1 depicts the proposed before, during and after phases of interaction with example trajectories in each phase.

Whether a map is useful for identifying and distinguishing interactions described by verbs depends on the features and relations that make up the map's axes. [reference removed] proposes that the *before* and *after* phases should map relative velocity against the 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 is the one doing the pushing, and the target is the body being pushed. By convention, the actor is designated as body A and the target is body 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 the measure of the 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 something 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. In [reference removed]'s original proposal, the second dimension of the during/contact map was a measure of the distance traveled by both bodies away from the initial contact-point. In the pilot studies we describe in the next section, we instead measured the amount of time the bodies were 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. Using these labels, an interaction can be described as an ordered triple of trajectory labels, representing the before, during and after characteristic trajectory.

For example, $\langle \mathbf{b}, \mathbf{b}, \mathbf{b} \rangle$ describes a *shove*. In a shove interaction, 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 it's velocity while the target moves at a greater velocity due to the energy imparted it (after **b**.

In Figure 2(b), below, we provide a plot of the dynamics of a simulated shove action (in the next section we describe the simulator we used to generate this action). The map in the figure plots the dynamics for a portion of the time between contact phases of repeated. 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 is about to take place..

Using 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 (after phase **d**). Harass highlights that all interactions are not to be viewed only as single movement to contact, but may involve many such movements to contact, one after another, and may even switch between different kinds of contact interactions.

- *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.

In the past year we conducted two exploratory studies that provided evidence that child and adult verb usage is influenced by the perception of dynamical features. We first describe the stimuli and procedures and then review the results from the studies.

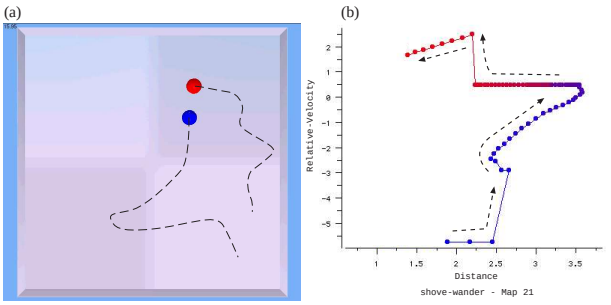

Figure 2: (a) Example of maps-for-verbs simulation running the shove-wander action, as rendered in breve. (Note: dashed lines represent motions of colored patches for demonstration purposes; only the moving color-patches themselves were displayed in the stimuli movies.); (b) Dynamic map plot of shove-wander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

Experiments

Stimuli

In both studies, 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" 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" used, 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. In the first study, children's responses were recorded and later transcribed, while in the second study, adults entered their responses directly into the computer using the computer keyboard.

Study 1 - Children

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 one time — with the exception of chase, which the child watched twice; chase appeared once in the first 9 trials, and then again in the second nine trials. 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.

Study 2 - Adults

Participants Forty-four undergraduates (M = 20.5 years old) at the University of Massachusetts participated in this study.

Procedure As with Study 1, a total of 18 movies were presented to each participant, with chase being viewed twice. After watching a movie, participants were asked to write down an answer to the questions on a sheet of paper given to them by the experimenter. The questions were the same for every movie:
1. What are the balls doing in this movie? (Give your overall impression of what was happening between them, the gist)
2. What is the red ball doing?
3. What is the blue ball doing?
4. Can you think of any words to describe the tone or the mood of the movie? (e.g., the balls are friendly/ not friendly)
The experimenter encouraged participants to write as much as they could to describe the movies.

	counter-shove	bounce	harass	hit	shove	push
chase	27.559	**56.300**	**58.034**	**58.038**	**62.051**	**61.025**
push	*46.174*	**50.019**	34.526	32.338	**61.974**	–
shove	41.365	*49.012*	40.026	34.300	–	–
hit	33.352	32.682	27.161	–	–	–
harass	27.953	29.427	–	–	–	–
bounce	30.013	–	–	–	–	–

Table 1: χ^2 scores for pairings of action-scenarios. **Bold** indicates score is significant, *italic* indicates marginal significance

Measures

In both studies, all the action words and other 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 unique 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.

After canonicalization, we kept only the verbs from the content words. As an example, the remaining canonical words from the study with children included the following 32 words: about, around, away, bonking, bouncing, bumping, catching, chasing, circle, coming, down, fast, flying, following, friends, getting, hitting, knocking, moving, not-moving, playing, pushing, running, slow, standing, staying, stopping, tag, together, trying, up, zooming. A similar list of 155 canonicalized content words was extracted from the adult responses.

Results

For the two pilot studies, we were interested in whether the words used by subjects to describe the movies would depend on the dynamical features present in the interactions. We therefore focused our analysis on the frequency of words used to describe each movie.

In Study 1 we conducted a series of χ^2 tests, summarized in Table 1, to determine if word usage was significantly different amongst different movies. We combined word frequencies for each of the variations (i.e., stationary, wander, coast) of each of the interaction types. Thus, push included all of the word frequencies for push-stationary, push-wander and push-coast. We then compared the interaction groups with each other. The .05 critical value for a χ^2 table with 31 degrees of freedom is roughly 45, so as can be seen from the table, many of the comparisons were not significant. However, χ^2 scores revealed that the words used for chase were significantly different from those used to describe the other interaction types, and push was significantly different from shove and harass (but not hit). Another χ^2 test comparing chase to everything else yielded a score of 92.266, which is highly significant. Another comparison, between push and everything else except chase yielded a score of 64.725, also significant. Finally, chase compared to the variations of push yielded a significant χ^2 score of 61.025.

These results suggest three general categories of interactions based on word usage: push, chase and interactions involving punctuated impact, such as shove and harass. While these results are suggestive and exciting, we believe there are a number of factors that contribute to the results not being stronger. Children are challenging to work with, especially in tasks requiring verbal response. Their vocabularies are small – we observed that the children in the study tended to use only a few verbs to describe the movies. For example, the word "bonk" was used by multiple children to lump together the movies for hit, harass and shove. Post-hoc analysis also revealed that children frequently used negatives (e.g., "not hitting") when describing the movies. Again, this is likely a function of their smaller vocabularies. In future studies we will want to code these as representing a single canonical form: e.g., `not-hitting`. Finally, the number of children in our study was relatively small. The structuring effects of the perception of dynamics on verb use is likely more subtle in children, so we expect we need a larger subject pool to draw out these effects. Our goal is to return to study children so that we can explore the development of the the relation between interaction dynamics and verb use. Before doing that, however, we conducted a study with adults – adult vocabulary is much more developed and consistent, and we expected to find the effects of dynamics on verb use more perspicuous.

As described in the previous section, the procedure and data collection in the adult study was basically the same as that used in the first study. However, we used hierarchical agglomerative clustering (hac) (Duda, Hart, & Stork 2001) to cluster the interaction classes based on word usage frequencies. This method produces trees representing the relative similarity of the items being clustered: items that are most similar are grouped first, and then those groups are merged with items or groups that are less similar, and so on. Each time two items or groups are merged, a parent node is added to the tree with the merged items as children. Eventually all items are grouped forming the root of the "tree." Figure 3 shows the generated dendrogram resulting from clustering the movies based on adult word usage frequencies. The leaves of the tree contain the names of the movies. Ignore for the moment the additional labels and notation to the right of the tree.

Looking at the dendrogram alone it is not clear how to interpret the groupings. However, recall that the movies were generated by behavioral programs written with the purpose of behaving according to the maps for verbs dynamical features (Figure 1). We are aware that the programs producing the interactions are not precise replications of the maps for verbs dynamics and part of our planned future work is to develop methods to objectively quantify the presence of dy-

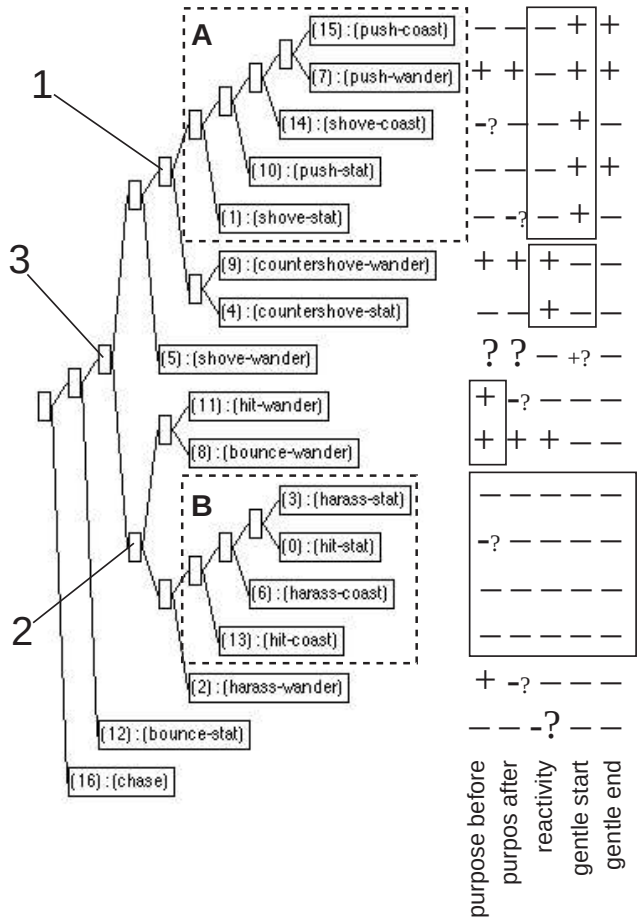

Figure 3: Dendrogram representing clustering of movies based on word usage frequencies, where word usage is based on the number of different subjects who used a given word. The complete set of 155 verbs was used to characterize word usage. The labels inside the leaves of the dendrogram correspond to movie names; the numbers are unique identifiers assigned by the clustering procedure and should be ignored.

namical features.

For the adult study, we approximated measurement of the dynamical features using human judges. Two adults that did not take part in the study were tasked with deriving a small set of features that would serve to distinguish each movie. The judges proposed five features: (1) whether target looked purposeful before or after contact (*purpose-before*, *purpose-after*) — "purposeful" was in terms of whether the target appeared to change its heading on its own; (2) whether the target seemed to react to contact (*reactive-during*) "react" was in terms of whether the target appeared to change its behavior based on blue's contact; and (3) whether the initial or final stages of the contact appeared gentle (*gentle-start*, *gentle-end*).

For each movie we assessed a minus (= no) or plus (= yes) depending on whether a given feature was present. For some movies we were uncertain about the presence of the feature, so we assigned a +? or -?; in cases where we were completely uncertain we gave a ?. The vector describing the presence or absence of features is pictured in Figure 3 on right-hand side of each leaf-node movie-label.

By associating the feature vectors with the dendrogram clusters, we now see that there are clear relations between features and clusters. The internal node labeled 1 in the dendrogram tree of Figure 3 appears to distinguish between (a) the group of movies for which the target is not reactive to the actors contact and for which the contact begins gently from (b) a group of movies for which the target is reactive and contact does not begin gently. The node labeled 2 in the dendrogram appears to distinguish between whether the target looks purposeful before or after interaction (although the placement of harass-wander is problematic it should be associated with hit-wander and bounce-wander). Finally, the node labeled 3 appears to separate (a) groups of movies that involve gentle starts to interactions or for which the target is reactive from (b) movies that all involve abrupt beginnings and endings to the contact phase of interaction (except for bounce-wander). We generated a second dendrogram using a more restricted set of verbs (verbs used by 10 or more subjects). The resulting tree was very similar, with the subtrees labeled A and B in Figure 3 appearing unchanged in both.

Discussion

These results, although at this stage only suggestive, are interesting. The dynamical features chosen by the human judges are associated with groupings of movies based on word usage. Also, our experience in collecting and analyzing the pilot study data has taught us that combining clustering based on word frequency with measurements of the presence of dynamic features uncovers how verb usage is influenced by dynamics. Our next set of planned studies will again look at adults and children, but address the following two issues.

First, instead of relying on human judges to assess the dynamic features of movies, such as gentle or violent contact and gradual or extreme acceleration, we will extract features automatically from movies, using algorithms we invented for finding structure in time series. As the movies each involve just two blobs moving on a two-dimensional plane, all the information in the movies is contained in time series of the x and y location of each blob (four series in all). It is a simple matter to detect patterns in data like these with, say, root mean squared (RMS) difference measures: A sequence that represents a pattern is passed over the raw time series or its derivative, and squared differences between expected (pattern) and observed (time series) data are computed. Low scores indicate places in the series that match a pattern well. [reference removed] compared movies to each other in this way. A limitation of this method is that we sometimes see variations in the duration of a pattern. For example, short and long "push" episodes are both pushes, but would not match well according to the RMS method. In this case we can use dynamic time warping or other edit-distances, as we have in the past with robot data [reference removed].

A more ambitious approach is to discover, not merely detect, dynamic features in movies. [reference removed] de-

veloped a method to learn new features in real-valued time series by modeling the series as a hidden markov model. Hidden states represent unknown processes that generate patterns in series. [reference removed] showed how to learn the parameters of these processes. Related work by [reference removed] models time series as markov chains and shows how to find distinctive clusters of subsequences of series, i.e., distinctive patterns in longer series. Finally, [reference removed] provides an unsupervised algorithm for discovering features in time series.

With these methods, we will be able to assign automatically to each movie a vector of dynamical features, or rather, a multinomial distribution of the frequency of occurrence of each feature in the movie. No human judgment is involved: featural characterizations of movies are objective, derived algorithmically and directly from the data in the movies.

The next question is whether these dynamical features correlate with verb choice. Recall that each movie is described by a frequency distribution over the content words used by subjects to describe it. For instance, the probability of "chase" being used to describe one movie might be 0.28, while "bump" has a very low probability; these proportions may be reversed for another movie. Thus, each movie is represented by a multinomial distribution over content words as well as a multinomial distribution of dynamical features. We can treat one distribution as "predictor" and another as response variables and run a multivariate regression to find the degree of joint linear dependence between these distributions. Randomization testing is an easy nonparametric way to find the probability of an observed degree of dependence under the null hypothesis that dynamic features are uncorrelated with verb choice [reference removed].

References

Barsalou, L. W. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences* 22:577–609.

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

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

Duda, R. O.; Hart, P. E.; and Stork, D. G. 2001. *Pattern Classification*. New York, NY: Wiley.

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

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

Johnson, M. 1987. *The Body in the Mind*. University of Chicago Press.

Klein, J. 2002. 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*. http://www.spiderland.org/breve/.

Lakoff, G. 1987. *Women, Fire and Dangerous Things*. University of Chicago Press.

Mandler, J. M. 1992. How to build a baby: Ii. conceptual primitives. *Psychological Review* 99(4):587–604.

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