When Push Comes to Shove: A Preliminary Study of the Relation Between Interaction Dynamics and Young Children's Verb Use

Clayton T. Morrison† (clayton@cs.umass.edu)
Erin N. Cannon‡ (ecannon@psych.umass.edu)
Paul R. Cohen† (cohen@cs.umass.edu)
Richard S. Bogartz‡ (bogartz@psych.umass.edu)
Carole R. Beal‡ (cbeal@psych.umass.edu)

†Department of Computer Science; University of Massachusetts
‡Department of Psychology; University of Massachusetts
Amherst, MA 01003 USA

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 the interaction. We report an initial investigation of this claim in a study in which young children were asked to describe movies of simple interactions governed by the dynamics proposed in the maps-for-verbs framework. The results are suggestive and motivate further investigation. We discuss what we learned and our plans for future study.

Introduction

A significant portion of our language is devoted to referring to, expressing, and representing the temporally extended dynamics of our world. Following Tomasello (1992), we label the class of words used to refer to such dynamics as "verbs." Cohen (1998) presents a framework for distinguishing interactions involving whole-body objects by considering the dynamics of the before, during (contact) and after phases of interaction. These phases are characterized using dynamic maps to plot various measures of the physical interaction. The *maps-for-verbs* framework proposes to use dynamic maps to represent the interactions referred to by certain classes of verbs. Here we report a preliminary study of the use of this framework to predict the verb use of young children describing simple whole-body interactions.

The framework advances two hypotheses:

Hypothesis 1 proposes that these representations (dynamic maps of before, during and after interaction) are a foundation for the semantics of verbs describing physical interactions between objects.

Hypothesis 2 proposes that the triptych representation of dynamics is present early in humans and therefore predicts that children make roughly the same distinctions in their early verb use.

We will have evidence for the hypotheses if features of actions in the maps-for-verbs framework are systematically associated with the frequencies with which verbs are used by children.

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, Cohen (1998), Cohen & Oates (1998), Rosenstein (1998) and Rosenstein, Cohen, Schmill, & Atkin (1997) have explored the use of dynamic maps as a suitable representation of activities. This work developed a set of techniques, adopted from non-linear 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

The maps-for-verbs framework proposes that simple interactions between whole bodies can be characterized by the physical dynamics of the interaction. The framework proposes that whole-body interactions are naturally divided into three phases: before, during and after contact. Figure 1 depicts

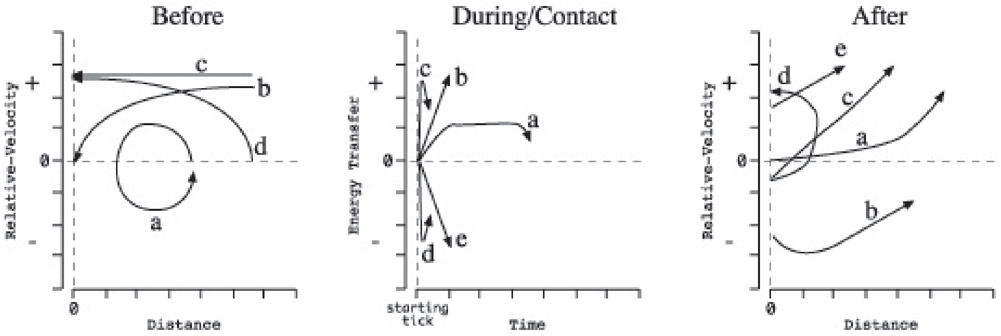

Figure 1: Maps-for-verbs model of the three phases of interaction.

these three phases. The interaction of two bodies can be plotted in a two-dimensional space called a *map* (also called a phase portrait or phase diagram). These maps portray the changes in features of or relationships between the two bodies, over time. A given interaction is then described as a trajectory through the map's dynamics space (example trajectories are shown in Figure 1). These maps enable identification of characteristic patterns present in the dynamics of classes of interactions.

Of course, whether the map tells us something useful about the dynamics of the interaction as it relates to identifying or modeling the referent of a verb depends on the features and relations that make up the map's dimensions. Cohen (1998) 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 scenario involving a PUSH, 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 *during* phase is proposed to be a map between perceived energy-transfer (from the actor to the target), and some other measure. 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: the target is imparting more energy to the actor. To measure perceived energy-transfer, we used the simplification of calculating the acceleration of the actor in the direction of the target while in contact. In the 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. For this work, we instead measured the amount of time the bodies were in contact.

To illustrate the use of these maps, consider the following seven interactions between whole bodies, described by the verbs push, shove, hit, harass, bounce, counter-shove and chase. Figure 1 depicts a set of labeled trajectories that characterize the component phases of these seven interaction types. Using these labels, an interaction can be described as a triple of trajectory labels, indicating the before during and after characteristic trajectory; for example, (**b**,**a**,**a**), which happens to describe a push. In a *push* 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 smoothly imparts more energy to the target, as illustrated by trajectory **a** of the during/contact phase. After breaking-off contact with the target, the agent then gradually increases its velocity and moves away from the target, illustrated by trajectory **a** in the after phase.

A *shove* starts out like a push, except that the contact phase involves a rapid increase in energy-transfer from the actor to the target (during **b**), and the after phase typically involves the actor rapidly decreasing it's velocity, while the target moves at a greater velocity from the energy imparted to it (after **b**). So the shove triple is (**b**,**b**,**b**). A demonstration of the dynamics for shove, based on our simulator (described below), is depicted in Figure 2(b). This map plots the dynamics for a portion of the time between contact phases, beginning with very low relative velocity, as would be expected just after completing the contact phase of a shove (after phase), and ending with a high relative velocity that is ramping down, as is expected in the period during which a new shove is about to take place, as the actor slows to prepare for contact (before phase).

Hit may begin with the actor already at high velocity relative to the target, or involve an increase in relative velocity (before **c** or **d**). In either case, the actor is moving toward the target. The actor then contacts the target at a high velocity, rapidly imparting a large amount of energy, for a brief amount of time (during **c**). And the denouement of the interaction leaves the actor at a slightly lower relative velocity (due to loss of momentum from striking the target): after **c**. The hit triple is then (**c**/**d**,**c**,**c**).

Harass is similar to a hit, except that the after-phase involves the actor quickly recovering its speed and moving back toward the target (after **d**), not allowing the distance between to two to get very large: (**c**/**d**,**c**,**d**). This cycle of interaction occurs repeatedly over a short amount of time. 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 and counter-shove are different from the above in that the target makes a more active response to the actor's actions. A *bounce* begins like a hit or harass, but at contact the target transfers a large amount of energy back to the actor (during **d**), making it look like a hit initiated by the target. The actor's velocity is then high relative to the target's in the after-phase (after **e**), resulting in the interaction triple (**c**/**d**,**d**,**e**). *Counter-shove* is the target's version of a shove, having a before phase like a push or hit, a contact phase with a sharp but more temporally extended increase in energy transfered *to* the actor, and an after phase like a bounce: (**b**/**c**/**d**,**e**,**e**).

Finally, *chase* involves the agent moving toward the target, closing the distance between the two, but never quite contacting the target, all while the target moves in a direction away from the agent. This is a special case of interaction in which there is likely no contact. This is depicted as the circular trajectory **a** in the before phase, in which relative velocity alternatively increases with a decrease in distance, followed by a decrease in relative velocity and an increase in distance. The interaction triple leaves the latter two phases blank: (**a**,-,-).

We used these seven classes of interaction as the basis for a study in which we looked at the frequency of verb usage of young children asked to describe the these interaction types after observing them.

Method

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. Written parental consent was obtained for each child. Participation was voluntary, and each child received stickers in return. Three of the sixteen children's data were dropped, one due to experimenter error, and two because of lack of response.

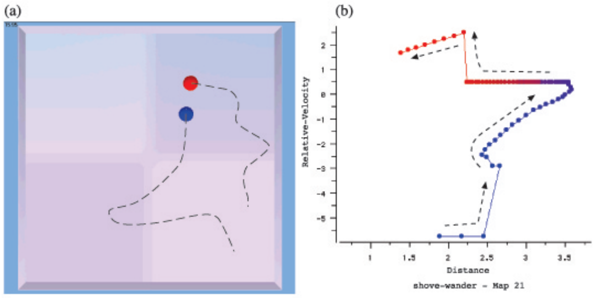

Figure 2: (a) Example of maps-for-verbs simulation running the shove-wander action, as rendered in breve. (Note: dashed lines represent motions of colored patches for demonstration purposes; only the moving color-patches themselves were displayed in the stimuli movies.); (b) Dynamic map plot of shove-wander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

Stimuli and Equipment

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 model is rendered as two generic objects (a blue ball for the actor and a red ball for the target) moving on a white platform (see Figure 2(a)).

We generated a set of movies based on the rendered interactions. 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.[1]

The 17 movies were recorded and presented in Mac OS X Quick Time Player 6.0.1, averaging

[1] 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.

	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

17.71 seconds in length (ranging from 15–20 seconds). Movies were presented on a Macintosh G3 iBook with a 14” screen. Children's responses were recorded using a Sony microcassette recorder.

Procedure

Testing took place in a day care center, in a small private room relatively free of typical distractions. Each child was randomly assigned one of two presentation orders for the movies. A total of 18 movies were presented to the child, 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. During presentation of the movie, if the child did give an answer, the experimenter followed up with the questions “Do you think they are friends?” then “Why?” or “Why not?”; these were usually effective in eliciting more verb usages. If the child wanted to see the movie again before giving an answer, or if the experimenter did not feel the movie was being attended to, the child could see it again, for a maximum of three times for each movie. Each child was tested in a minimum of two sessions to complete all 18 trials.

Measures

Each participant's response was recorded and later transcribed by an experimenter. Then 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 only kept the content words that occurred three times or more, leaving 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.

Results

Quantitative

We conducted a series of χ^2 tests, summarized in Table 1. Each row and column label represents the combination of all variations (stationary, wander, coast – see footnote 1) of the labeled interaction type (thus “push” includes push-stationary, push-wander and push-coast). For each comparison, we built a contingency table crossing frequency of each of the selected content words appearing in a given interaction-type against the two indicated interaction-type groups. 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, chase is significantly different from just about everything else, and push is different from shove and harass (but not hit).

Using the same kind of contingency table definition as above, 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.

Qualitative

For a different perspective, we used hierarchical agglomerative clustering (hac) (Duda, Hart & Stork, 2001) to cluster the interaction classes based on word frequencies. We used a measure of similarity based on *ranked* frequencies, as follows. For each interaction class we created a vector representing the frequency of each content word (that is, the number of times that word appeared in transcripts associated with that class). A rank was then assigned to each vector-element value, such that the most frequent word gets a rank of 0, the second-most a rank of 1, etc. The rank frequency measure of a comparison between two such vectors is then the absolute value of the differences in ranks. This measure results in

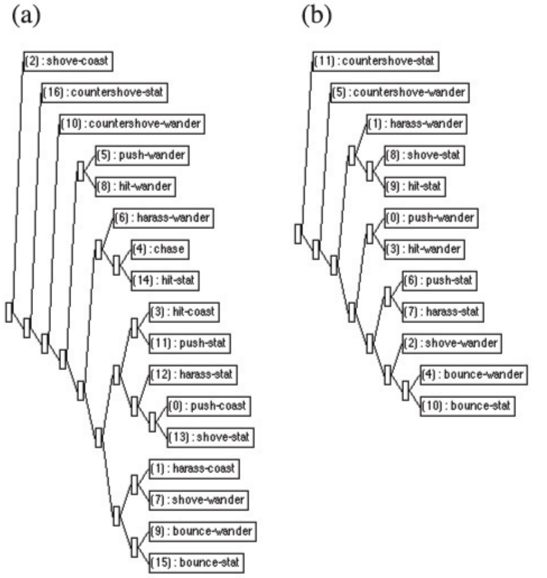

Figure 3: Dendrograms generated from hierarchical agglomerative clustering of interaction classes based on rank distance measure. (a) Dendrogram for all interaction classes except for chase. (b) Dendrogram for only wandering and stationary target interaction classes

lower scores for vectors in which higher frequency words are shared. Used as a distance measure in hac, this in turn means that such vectors will tend to be clustered earlier, indicating their similarity with respect to word frequencies.

Dendrograms representing the hierarchical structure of hac are depicted in Figure 3. Figure 3(a) shows the dendrogram that results from clustering all interaction classes. It is difficult to see much structure in this dendrogram. However, we noticed that the coast versions of the interactions were distributed amongst the subclusters. We repeated the hac procedure using only the classes with wander or stationary versions, resulting in the dendrogram of Figure 3(b). This dendrogram is interesting because both bounce variations, as well as both countershoves, appear related, and except for harass-wander and shove-wander, the other subcluster pairs involve both stationary or both wander variations of interaction classes.

Discussion

While the effects are not overwhelming, we have found preliminary evidence that children watching the movies of interactions based on the maps-for-verbs parameters are systematically selecting particular words in response to the dynamics of the movies. The χ^2 scores reported in Table 1 suggest three general categories of interactions, based on word usage: push, chase and interactions involving punctuated impact, such as shove and harass. Additionally, the hac dendrograms suggest there is structuring of word usage based on the behavior of the target in the movies.

Where does this leave us? At this point, the status of hypotheses 1 and 2 is still uncertain. We have demonstrated that words are selected preferentially in response to different dynamics, but we have not demonstrated that the distinctions in the maps for verbs framework (i.e., the different paths through the three phases) are systematically associated with different distributions of evoked words. Word use certainly seems to be associated with dynamics, but not necessarily in the ways described by the maps-for-verbs framwework. More work is needed to show that the framework *predicts* distributions of word use for different movies.

We do, however, have enough evidence to warrant further investigation, and we have learned a number of important lessons from this preliminary study. The following are some issues we have identified that we will want to control for.

(1) Based on additional analysis of the data, we found what may be a practice affect in the children's responses. For example, if they use the word "bounce" for one movie, then they tended to use that word again for subsequent movies. One possible explanation for this is that children at this age have underdeveloped vocabularies. In future studies of this age group we will attempt to asses the child's vocabulary and work to control for any practice effects.

(2) In general, the data we do have is very sparse. Most of the content words are used very infrequently and often used in several action classes, making it unclear whether their use indicates structuring by dynamics or is meaningless. Again, small vocabularies and varying linguistic competence of the children in the study make clear that we need to collect more subject data to determine whether the apparent structure is real.

(3) From a different perspective, evidence from work in motion capture and animation (Bregler, 2003) suggests that humans perceive slightly exaggerated motion as more realistic than filmed real-life motion rendered in two dimensions (this is why the art of cartooning tends toward exaggerated motion). It is quite likely that some of the dynamics expressed in our movies are too subtle to be noticed.

Our experience from conducting this study suggests a clear path to improving our methods for future investigation. Based on the above observations we plan the following for the next phase of our investigation.

(1) We need to "Disneyfy" the interactions. We plan to create new movies with slightly exaggerated movements, but still following the parameters of the

dynamics in the maps-for-verbs model.

(2) Currently, there is still some distance between our specification of the dynamic behavior of the interacting colored balls and the dynamics predicted by the maps-for-verbs model. We are working to develop a method by which we can go *from* specified dynamic maps *to* interaction behavior. This way we could "draw" the dynamics (the relationships over time) of the interaction first (that is, specify a trajectory first), and then generate a movie of the interaction we specified. This would allow us to have strict control over the stimuli, generating it directly from the model.

(3) We will collect data from a range of human subject ages, including older children and adults.

(4) We would like to develop methods to measure the existing vocabularies of the age-group of children that participated in this study in order to better understand how linguistic competence might have affected their responses.

Acknowledgments

Thanks are due to Jon Klein for technical help with the breve environment.

References

Blythe, P. W., Todd, P. M. & Miller, G. F. (1999). How motion reveals intention. Categorizing social interactions. In G. Gigerenzer, P. M. Todd, & the ABC Research Group, *Simple heuristics that make us smart* (pp. 257–285). New York: Oxford University Press.

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

Bregler, C. (2003, Jan.). Turning to the masters: Motion capturing, computer animation, and cartoons. Paper presented at the *DIMI Workshop on Perceptive Social Agents and Robots*, La Jolla CA.

Cohen, Paul R. (1998). Maps for Verbs. *Proceedings of the Information and Technology Systems Conference, Fifteenth IFIP World Computer Congress* (pp. 21–33).

Cohen, Paul R. & Tim Oates. (1998). A Dynamical Basis for the Semantic Content of Verbs. In *Grounding of Word Meaning: Data & Models Workshop*, AAAI-98, pp. 5-8.

Duda, Richard O., Peter E. Hart & David G. Stork. (2001). *Pattern Classification.* New York, NY: Wiley.

Heider, F., & M. Simmel (1944). An Experimental Study of Apparent Behaviour. *American Journal of Psychology 57* (2) (pp. 243–59).

Intille, S. & A.. Bobick. A framework for recognizing multi-agent action from visual evidence. *Proceedings of the Sixteenth National Conference on Artificial Intelligence* (pp. 518-525). Orlando, Florida, July 1999.

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

Rosenstein, Michael T. (1998). Concepts From Time Series. *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (pp. 739–745).

Rosenstein, Michael T., Paul R. Cohen, Matthew D. Schmill & Marc S. Atkin. (1997). Action Representation, Prediction and Concepts. Paper presented at *AAAI Workshop on Robots, Softbots, Immobots: Theories of Action, Planning and Control.*

Todd, P.M., & Barrett, H.C. (2000). Judgment of domain-specific intentionality based solely on motion cues. Presented at the *12th Annual Meeting of the Human Behavior and Evolution Society* Amherst, MA.

Tomasello, Michael. (1992). *First verbs: a case study of early grammatical development.* Cambridge [England]; New York, NY, USA: Cambridge University Press.