Extending Embodied Lexical Development

David Bailey, Nancy Chang, Jerome Feldman, and Srini Narayanan
{dbailey, nchang, jfeldman, snarayan}@icsi.berkeley.edu
International Computer Science Institute and University of California, Berkeley
1947 Center Street Suite 600, Berkeley CA 94704 USA

Abstract

This paper describes an implemented computational model of lexical development for the case of action verbs. A simulated agent is trained by an informant labeling the agent's actions (here hand motions), and the system learns to both label and carry out similar actions. The verb learning model is placed in the broader context of the NTL project on embodied natural language and its acquisition. Based on experimental results and projections to the full range of early lexemes, a significantly enriched model is proposed and its properties discussed.

Introduction

The embodiment of concepts and language is a central issue in cognitive science. How can a neural system represent and learn concepts, and organize them into a set of lexical items? For almost a decade, the Neural Theory of Language (NTL) group at ICSI and UC Berkeley has sought computational insight into these questions by asking them of structured connectionist systems, rather than the physical neural systems of the brain (Feldman et al., 1996). The basic questions of neural and cognitive development are receiving increasing attention from the connectionist perspective. Although there is a good deal of theoretical and modeling work on the acquisition of syntax there does not appear to be any detailed theory of lexical development comparable to the NTL project. This paper is an overview some of of our recent results and challenges.

One critical empirical finding from studies of language acquisition is that the child's first words label not only things, but also relationships, actions and internal states (Tomasello, 1995). Clearly, embodiment is central to all of these. Early lexical development thus provides an ideal task for studying embodied cognition, since we can isolate linguistically and conceptually simple situations for which to construct and test detailed models.

Our first major effort was the dissertation work of Regier (1996), a computational model of how some lexical items describing spatial relations might develop in different languages. Since languages differ radically in how spatial relations are conceptualized, there was no obvious set of primitive features to build into the program. The key to Regier's success came directly from embodiment: all people have the same visual system from which all visual concepts must arise. By including a simple but realistic visual system model, Regier's program was able to learn spatial terms from labeled example movies for a wide range of languages, using conventional back-propagation techniques.

The project's scope was expanded to verbs with Bailey's dissertation work (Bailey, 1997; Bailey et al., 1997), a computational model that learns to produce verb labels for actions and also carry out actions specified by verbs that it has learned. A shortcoming of the standard view of lexical acquisition is that it provides no account of how a child learns to *make use* of the concepts she learns and the words that label them. This same weakness appears as a technical consequence of using back-propagation in Regier's work and in PDP models: even when the network learns perfectly how to classify a domain, it has no mechanism for executing the action. Bailey's work addresses this shortcoming by employing learning algorithms that produce usable representations of actions. After training on examples of action-word pairs, the system can produce an appropriate label for a particular motor action based on features of both the action and the world state. In addition, however, the learned verb representation also functions as a command interface that allows the system to execute a given verb.

Cross-linguistic experiments with both Regier's spatial relations network and Bailey's verb-learning system reveal both strengths and weaknesses of the current state of development. We believe that the basic principle that early word learning across languages can be modeled very well by embodiment-based structured connectionist models has been established. On the other hand, it is clear that our systems must incorporate much richer models of the neural substrate to handle even the early lexical development of children. To better estimate what is required, the group is beginning to study the full range of early word learning, rather than continuing to focus on isolated sub-vocabularies.

The original name of the project, L_0, was chosen because zero was the approximate percentage of language we were attempting to cover. The current effort is still concerned with only a tiny fraction of the complexity of language learning, but because we are now grappling with all of a child's first (say) 200 words, we have presumptuously renamed the project NTL. In this paper, we outline plans for expanding our detailed connectionist modeling to cover all early lexical acquisition. As always, the theories and systems are intended to apply to all natural languages.

Representational mechanisms

To bridge the gap from embodied experience to its expression as abstract symbols in language, we have found it necessary to work at multiple levels of description. Regier's work, for instance, linked the connectionist and cognitive levels, with the neural level implicit. Subsequent more complicated domains have required us to add a computational level as an abstraction from the connectionist level. Although the focus of this paper is this computational level, the NTL papers in the 1997 Conference of the Cognitive Science Society spanned all five levels:

cognitive:	words, concepts
computational:	f-structs, x-schemas (see below)
connectionist:	structured models, learning rules
computational neuroscience:	detailed neural models
neural:	[still implicit]

Our computational level is analogous to Marr's and comprises a mixture of familiar notions like feature structures and a novel representation, executing schemas, described below. Apart from providing a valuable scientific language for specifying proposed structures and mechanisms, these representational formalisms can be implemented in simulations to allow us to test our hypotheses. They also support computational learning algorithms so we can use them in experiments on acquisition. Importantly, these computational mechanisms are all reducible to structured connectionist models so that embodiment can be realized.

The most novel computational feature of our current effort is our representation of actions, **executing schemas** (**x-schemas** for short), so named to distinguish them from other notions of schema and to remind us that they are intended to execute when invoked. We represent x-schemas using an extension of a computational formalism known as Petri nets (Murata, 1989). A Petri net is a bipartite graph containing **places** (drawn as circles) and **transitions** (rectangles). Places hold **tokens** and represent predicates about the world state or internal state. Transitions are the active component. When all places pointing into a transition contain an adequate number of tokens (usually 1), the transition is enabled and may fire, removing its input tokens and depositing a new set of tokens in its output places. X-schemas cleanly capture sequentiality, concurrency and event-based asynchronous control; with our extensions they also model hierarchy and parameterization.

To keep things minimal, our models use only one other computational mechanism—**feature structures** (**f-structs** for short, drawn as a row of double-boxes). F-structs are used for static knowledge representation, parameter setting, and binding. They have been chosen to be compatible with the "f-structures" in the literature on unification grammars, and are similar to well-known AI slot-filler mechanisms. From these simple constructs, a wide variety of modeling structures can be built.

The bottom third of Figure 1 depicts an example x-schema for sliding an object on a tabletop. The SLIDE x-schema captures the fact that people shape the hand while moving the arm to an object and that large and small objects are handled differently. It includes a loop that continues motion when not yet at the goal and a separate little schema for tightening the grip if slip is detected.

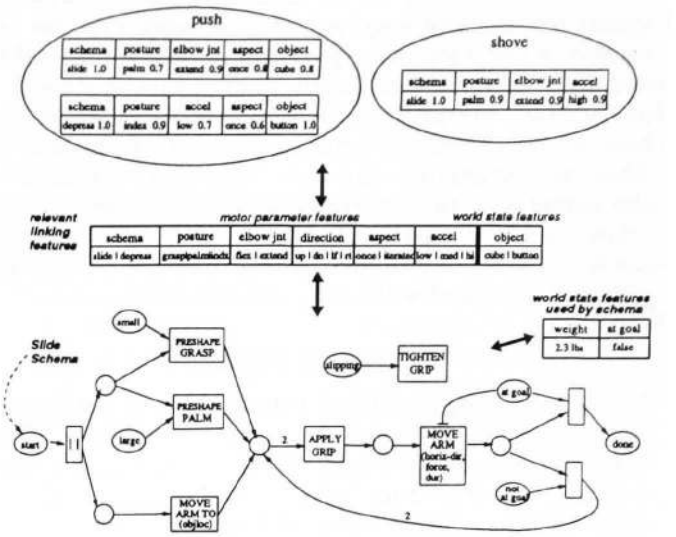

Figure 1: An overview of the verb-learner at the computational level, showing details of the SLIDE x-schema, some linking features, and two verbs: *push* (with two senses) and *shove* (with one sense).

Bailey's verb learning model

An overview of Bailey's verb learning system is given in Figure 1. In this model a special **linking f-struct** (center of Figure 1) plays an important role as the sole interface between language and action. It maintains bidirectional connections to the x-schemas: an x-schema receives bindings from f-structs and produces additional bindings during its execution. In this way, actions can be translated to and from semantic features. More generally, we claim that the requirements of parameterizing x-schemas are the principal determiner of which semantic features get encoded in a language. One critical linking feature is the name of the x-schema generating the action. Others include motor parameters such as force, elbow joint motion, and hand posture. Some world state features are also relevant, such as object shape.

Each sense of a verb is represented in the model by an f-struct whose feature values are probability distributions. Features are presumed independent and the representation is conjunctive or gestalt-like in nature. The top third of Figure 1 shows several word senses for the verbs *push* and *shove*. The upper left ellipse gives f-structs for two senses of *push*. The top sense is a hand motion that invokes the SLIDE x-schema. The ellipse on the upper right shows that *shove* also codes for the SLIDE x-schema but specifies high acceleration.

In execution mode, a verbal command is interpreted by choosing the sense that best matches the current

world state. This sense is in turn used to set the linking f-struct, thus determining which x-schema is to execute and with what parameters. For example, *shove* specifies both a SLIDE x-schema and high acceleration, but the force required depends (at least) on the size of the object involved, which is not specified in the utterance.

The verb learning model assumes that the child (or agent) has already acquired various x-schemas for the actions of one hand manipulating an object on a table, and that an informant labels actions that the agent is performing. As in Regier's work, we avoid some hard but, we feel, separable issues by assuming that the informant supplies just the verb. The problem faced by the model (and the child) is thus to learn how the verbs relate to its actions and goals. The detailed learning mechanisms are explained in (Bailey et al., 1997) and (Bailey, 1997) and will not be described here.

Learning results

Extensive testing of the verb learning system has demonstrated its ability to acquire some important distinctions between verbs of hand motion. For English, the system acquired 18 verbs from 200 labeled examples, with features such as schema name and hand posture playing a more important role in determining word sense than object size and direction of motion. The system had a 78% success rate for recognizing new examples. The relatively better performance of the system in obeying commands (81%) is not surprising, since it was evaluated by executing its learned model of the specified word and then trying to recognize the action. Interestingly, for both types of testing, errors consistently involved subtle distinctions, such as that between *heave* and *lift*, that caused the system to choose a plausible alternative. There were no gross errors.

Experiments in Farsi, Hebrew and Russian have confirmed the system's ability to model cross-linguistic variation, with many parameters used during training proving robust across experiments. Results echo those for English in some respects: most performance errors involved closely related verbs, with the system often favoring specific verbs over more general ones. But as expected, the lexicons acquired for each language differ significantly. For example, the distinction between Farsi *hol daadan* (away-directed motion) and *feshaar daadan* (applying force without motion) depends more on force and duration than that between English *push* and *pull*. The Hebrew lexicon was smaller than the English lexicon and involved more general verbs of motion, such as *hirzik* (make-far) and *kerev* (make-close); reflecting a typological tendency of verbs to encode either path (Hebrew) or manner (English). Results for this simplified Hebrew were even stronger than those for English.

Differing lexicalization patterns also shed light on weaknesses of the model. Some of these, such as the need for more world features (such as object weight), are easily remedied. Other problems suggest that some refinement of the underlying model may be necessary: in general, the system has not been designed to capture either correlations between features (such as that between object size and force for *push*) or abstract verb senses (such as *move*) needed to learn hierarchical lexicons.

Challenges of a rather different nature account for the system's difficulties in handling the widespread grammaticization of aspect and deixis in Russian verb affixes: although some aspectual distinctions were acquired from a simplified training set of Russian verbs, the system had difficulties acquiring word senses in which deictic reference depends on multiple inter-dependent features. These problems arise for English verb phrases as well: the direction of motion in *take away* is opposite that of *give away* despite their common particle, since they have different deictic centers. It also remains unclear whether the current system can handle more complicated problems involving aspectual composition and goal-oriented actions. These shortcomings underscore the need for a richer model of intentional state that can represent alternate perspectives and goal-driven actions.

The remainder of the paper outlines some extensions to the x-schema representational framework that address some of these problems and suggests how these additions should significantly extend the range of learnable words.

An x-schema simulation environment

In recent work (Narayanan, 1997), we have extended the basic x-schema representation to model domain theories, with the same mechanism used for acting and reasoning about actions in a dynamic environment. The basic idea is simple. We assume that people can execute x-schemas with respect to f-structs that are not linked to the body and the here and now. In this case, x-schema actions are not carried out directly but instead trigger simulations of an imagined situation. It is easy to imagine, for instance, sliding King Kong up to the top of the Empire State Building and to predict what happens upon letting go. The NTL model for such mechanical planning assumes that the SLIDE x-schema can run with respect to a world model with simple qualitative physics. Physical models that people normally use appear to be simple enough to fit our paradigm fairly well, and some elementary ones have been implemented. We model the physical world as independent x-schemas with links to the x-schema representing the planned action. These x-schemas can interrupt one another or otherwise affect execution, for instance by changing f-struct values. A simplified simulation of the dropping of an object and the corresponding world simulation, depicted in Figure 2, illustrates the central ideas.

On the left of Figure 2, we have an x-schema corresponding to the important control transitions of the underlying DROP x-schema (Narayanan, 1997). On the right are x-schemas corresponding to the agent's simulation of the world. Both the agent's actions and the world's simulated evolution affect the agent's mental state. At the start of the simulation, the object is supported by an agent who then withdraws support as as result of the DROP action, consuming the token at the place labeled supported(obj1) and, through an inhibitory link, triggers the FALL x-schema. FALL simulates the

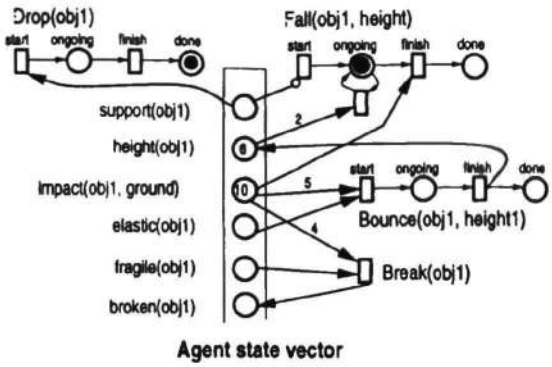

Figure 2: X-schema simulation of DROP, resulting in the object falling and either breaking or bouncing.

decreasing height of the object until it hits the ground.[1] Once the object hits the ground, it may bounce back to some new height, depending on the impact force and the object's elasticity. If the object is brittle (like a vase), it does not bounce but will instead break.

In our simulation framework, whenever an executing x-schema makes a control transition, it potentially modifies state, leading to asynchronous and parallel triggering or inhibition of other x-schemas. We believe such a system design supports a broad notion of action, in that the same active representation can be used for monitoring, control and inference. The notion of state as a graph marking is inherently distributed over the network, so the working memory of an x-schema-based inference system is distributed over the entire set of x-schemas and f-structs. Of course, this is intended to model the massively parallel computation of the brain.

This x-schema simulation framework has already proven useful in modeling metaphoric reasoning about event descriptions in abstract domains such as international economics. A crucial aspect of the implemented model is its capacity to exploit domain knowledge of spatial motion and manipulation (implemented as x-schema simulations) for real-time simulative inference. Results of applying our model to discourse fragments from newspaper stories in international economics show that crucial facts about abstract plans, goals, resources and intent can be expressed by projections from embodied concepts. Further details of this work can be found in (Narayanan, 1997). In the rest of this paper, we show how using and extending this simulation framework helps us overcome some of the limitations outlined earlier and cover a wider range of early child language.

Extensions

In this section, we examine the complete early child vocabulary and its implications for the direction of our modeling efforts. Although early vocabularies differ widely both within and across languages, published studies are sufficient for identifying the most common words. As is well known, the most frequent word types (although not tokens) are nouns, which we continue to view as relatively unproblematic cases to be dealt with later. [2] Similarly, early adjectives seem comparatively straightforward from a computational modeling perspective and moreover are not very common initially. Deferring work on nouns and adjectives allows us to focus on the more challenging early words.

For concreteness, let us consider the 49 words used by the most children in the Bloom (1993) preschool study. Half of these are nouns or names of sounds like *boom*, *moo* or *woof*. When we consider the other words, several issues come to the fore. In particular, two methodological issues come prior to any detailed modeling: First, it is known that children often use a word very differently from how an adult would. For example, *down* and *more* from the right hand column in Table 1 are often initially requests. This is closely related to the second, more general issue: any serious modeling must be based on the physical and intentional context in which the child is using a given word. The literature is not complete in this regard, but it is good enough to get us started.

box	cookie						
choo-choo	door						
get	eye	banana					
girl	go	boom					
hammer	here	bottle	apple				
horse	moo	cow	boy				baby
in	no more	daddy	that			mommy	ball
out	on	shoe	this	bead	bye	no	down
sit	truck	spoon	uhoh	open	hi	oh	juice
two	woof	there	whee	yes	yum	up	more
7	8	9	10	11	12	13	14

Number of children

Table 1: Words learned by 7 or more children, reproduced from Bloom (1993).

Several of the 23 words in the table that are not nouns or names of sounds (shown in bold) appear to be covered by our previous efforts to model the acquisition of spatial relation terms (Regier) or verbs of personal movement (Bailey), although really learning even these words involves some rather deep issues to which we will return later in this section. With this proviso, we can count four verbs of action (*get*, *sit*, *go* and *open*) and five spatial relation words (*in*, *out*, *on*, *up* and *down*) as understood.

Beyond the cases arguably covered by our previous models, the remaining lexemes again cluster into a small number of basic kinds. There are four that appear to be express emotion: *uhoh*, *whee*, *oh* (surprise) and *yum*. For our purposes, these present no basic problems. One assumes that the child may often hear a parent making the appropriate sound that correlates with the child's own emotional state, often with additional cues in the form of intonation patterns and facial expressions. Obviously enough, serious computational modeling of emotion is not to be taken lightly, but it doesn't seem necessary for our current purposes.

The meanings of the remaining items all depend on conversational context or reference. There are two greet-

[1] To simplify exposition, we leave out the relationship computed between the initial height and impact strength.

[2] Of course, to really implement an object naming system would require solving the computer vision problem.

ings, *hi* and *bye*, and four general communication terms, so called because their communicative content is understandable only relative to their discourse context; these are *yes*, *no*, *more* and *no more*. Also, there are two adult spatial adverbs, *here* and *there*, which again rely crucially for their meanings on the physical context of the conversation.

Both *this* and *that* similarly have referential potential only in conversational context. Note also that in adult speech they are each members of two closed classes: articles and pronouns. As such they play important functions at both the discourse level (by pointing anaphorically to other discourse elements) and the sentence-syntactic level (since articles are associated with an extremely restricted range of syntactic environments and thus are good predictors of the syntactic class of the next item). Beyond their semantic dependence on conversational setting, then, the development of specific grammatical functions for these lexical items poses a particular challenge and opportunity for an embodied NTL model of how they might be learned.

As this sampling demonstrates, many early lexemes depend on conversational context and thus present the same basic challenge to the NTL paradigm: modeling other people.We can simplify this task by assuming that these lexemes originally label or augment pre-verbal communication acts. Children develop communication patterns with their parents well before they speak; there are patterns of shared eye movements and other physical and vocal gestures (Foster, 1990) that form an obvious possible substrate for the deictic articles *this* and *that*. And no parent needs to be told about *no*.

Under the assumption that pre-linguistic communication routines are a necessary precursor to learning communication terms, the challenge to the NTL paradigm becomes one of modeling the interaction of the child's own mental representations with its representations of other agents. This simulation of other agents is a direct extension of the general x-schema simulation environment described earlier. Instead of passive x-schemas representing the physical world as in Figure 2, we use x-schemas that model the other agent. The interactions between the model's plans and the anticipated response of the other agent are computationally the same as in the case of physical simulation, although models of other agents must be much richer, including models of both their actions and their beliefs.

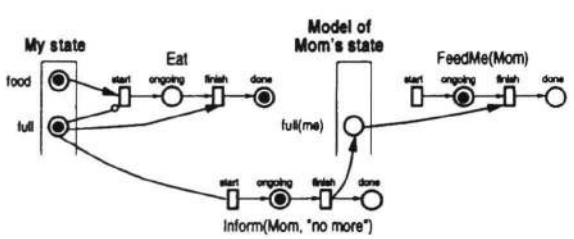

Figure 3: Depiction of child's own state and model of her mother's state in a context in which the child might say *no more* to halt feeding.

Figure 3 illustrates one context in which the phrase *no more* might be used. The simplified x-schemas and state representations depict both the child's own actions (EAT, INFORM) and her model of her mother's FEEDME action. The EAT schema on the left is enabled by the presence of food and continues until the child is full. The figure models a point when the child, having eaten her fill, is using the phrase *no more* to informing her mother that she is full, which she knows will stop her mother's feeding process. This communicative act presumably replaces or augments earlier gestural x-schemas.

This approach accommodates many simple cases in which a lexeme labels some action in the child's mental model of (conversational) context. But more complicated semantic distinctions will require additional mechanisms. Let us return to the four action verbs and five spatial relation words that appear to be explained by previous NTL models. The previous models took a fixed perspective with respect to which all words were learned. Regier's spatial term network takes what we call the **observer** perspective; the learning agent views scenarios that are then labeled. Bailey's action verb assumes what we call the **agent** perspective; the learning agent receives labels for its own actions. Regier gave no hint how an agent who learns to label a scene as *into* would know how to apply this label to its own actions. Bailey is similarly silent on how an agent who learns that one of its actions can be called *push* would recognize the same action when carried out by others. The situation is even more complex than this — there is at least one additional basic perspective, which we call the **experiencer**. Being pushed is experientially quite distinct from either pushing or observing some third party pushing. Similarly, it is quite different to put a toy *in* your mouth, see milk put *in* the refrigerator or be put *in* your bath.

As is clear from the fifty words of our sample, children's initial word meanings may take any of the three basic perspectives of agent, experiencer or observer. Most nouns are, of course, learned by observation, although some (*eye*, *boy*) might be learned first as part of one's own body or as a reference to oneself. Emotion words and actions like *sit* (and Bailey's examples) are normally learned first from the agent perspective. At least for American middle class children, words like *up* and *down* are first used in the experiencer perspective — the child is picked up or put down when the word is used. It isn't terribly important which perspective comes first for some lexeme for a given child — the question is how all of these come to be associated with the same term. Crucially, can a term learned from one perspective be understood and used from the others? We know of no systematic study of this transfer, but the anecdotal evidence suggests that it is common.

The multiple perspectives and the apparent ease with which children transfer among them presents a strong challenge for the NTL project of constructing detailed, neurally plausible models. Our proposed solution is to further extend x-schemas to support recognition as well as execution. Referring once again to Figure 1, one can imagine that an agent has the ability to recognize some-

one else carrying out a SLIDE action. That is, the x-schema formalism can also be used as an active template to recognize actions as well as carry them out.

This solution is computationally elegant and has some experimental basis, but it is still very speculative. In fact, there have been connectionist models that work exactly this way. The most relevant is Goddard's thesis system (1992) that recognizes human gaits from stick figure movie input. Goddard found that the best way to recognize a motion was to have an x-schema-like active representation that was brought into synchronization with the incoming visual data. In a good match, the simulation predicted the input stream and the visual recognition became easy. Alternative models competed in the usual connectionist way to provide the best match for a data stream.

There is also some developmental and biological support for this kind of model, often discussed under the rubric of imitation. Despite controversy about the extent to which other animals share this ability (Hauser, 1996), imitation is clearly a crucial aspect of human learning. Children can imitate a limited range of facial expressions just after birth, suggesting a connection between recognizing one of these behaviors and the motor schema for carrying it out. The animal literature contains reports of cells in monkey pre-frontal cortex that fire actively either when the monkey itself carries out a specific action or when it sees another primate carry out a similar action (Gallese et al., 1996).

The idea of using x-schemas for execution, inference and recognition might also help with another major shortcoming of Bailey's verb learning paradigm. Bailey's system learns only the most concrete embodiment of a word like *push*. But there is a more general, abstract meaning as well. This might be glossed as moving an object away from a deictic center using force directed through the object. Much of the cognitive linguistics literature is concerned with these general image-schematic (Lakoff, 1987) and force-dynamic (Talmy, 1988) semantic representations. In our Bloom data, three of the four verbs of action are the general forms: *get*, *go* and *open*, but these might refer to specific actions; *get* might refer to a pulling action representable directly in Figure 1. While it is still not known whether children develop the general meanings early, the model must allow for the possibility that they do.

Our current idea is to allow the embodied semantics for early action words to have both specific and general components. In this formulation, the SLIDE x-schema of Figure 1 would be accompanied by a general x-schema for achieving the goal of moving a physical object to a desired place. In learning *push* the child might associate the word with either the specific action, the general goal achievement or both. Recent work by C. Johnson (1997) suggests that early word learning might conflate specific and general meanings, such as *view* and *know* for *see*. This early conflation may serve as the basis for later metaphorical mappings in a manner that fits very well with the NTL paradigm. By applying recognition to these general x-schemas, we may have the basis for the inference of the goals and intentions of other agents, a crucial step much studied in AI.

Acknowledgments

This work was supported in part by NSF grant IRI-9619293 and a NSF Graduate Fellowship to Nancy Chang. Thanks to Masha Brodsky (for Russian experiments) and Nurit Melnik (for Hebrew experiments). Claudia Brugman, George Lakoff and the NTL group provided valuable comments on the paper and its underlying ideas.

References

Bailey, D. R. (1997). *When Push Comes to Shove: A Computational Model of the Role of Motor Control in the Acquisition of Action Verbs.* PhD thesis, Computer Science Division, EECS Department, University of California at Berkeley.

Bailey, D. R., Feldman, J. A., Narayanan, S., and Lakoff, G. (1997). Modeling embodied lexical development. In *Proceedings of the 19th Cognitive Science Society Conference.*

Feldman, J. A., Lakoff, G., Bailey, D. R., Narayanan, S., Regier, T., and Stolcke, A. (1996). L_0—the first five years of an automated language acquisition project. *AI Review*, 10:103–129. Special issue on Integration of Natural Language and Vision Processing.

Foster, S. (1990). *The communicative competence of young children : a modular approach.* Longman.

Gallese, V., Fadiga, L., Fogassi, L., and Rizzolati, G. (1996). Action recognition in the premotor cortex.

Goddard, N. (1992). *The Perception of Articulated Motion: Recognizing Moving Light Displays.* PhD thesis, University of Rochester.

Hauser, M. D. (1996). *The evolution of communication.* MIT Press, Cambridge, MA.

Johnson, C. (1997). Learnability and the acquisition of multiple senses: Source reconsidered. In *Proceedings of the 22nd Annual Meeting of the Berkeley Linguistics Society.*

Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind.* University of Chicago Press.

Murata, T. (1989). Petri nets: Properties, analysis and applications. *Proceedings of IEEE*, 77(4):541–580.

Narayanan, S. (1997). *Knowledge-based Action Representations for Metaphor and Aspect (KARMA).* PhD thesis, Computer Science Division, EECS Department, University of California at Berkeley.

Regier, T. (1996). *The Human Semantic Potential.* MIT Press.

Talmy, L. (1988). Force dynamics in language and thought. *Cognitive Science*, 12:49–100.

Tomasello, M., editors (1995). *Beyond Names for Things: Young Children's Acquisition of Verbs.* Lawrence Erlbaum Associates, Hillsdale, NJ.