L_0—The First Four Years

J. Feldman G. Lakoff D. Bailey
S. Narayanan T. Regier A. Stolcke

December 1993

Abstract

A summary of the progress and plans of the L_0 project as of December 1993. To be published in AI Review, Vol. 8, special issue on Integration of Natural Language and Vision Processing, edited by Paul McKevitt.

1 Introduction

The L_0 project at ICSI and UC Berkeley attempts to combine not only vision and natural language modelling, but also learning. The original task was put forward in (Feldman *et al.* 1990a) as a touchstone task for AI and cognitive science in a very simple from:

> The system is given examples of pictures paired with true statements about those pictures in an arbitrary natural language.
>
> The system is to learn the relevant portion of the language well enough so that given a novel sentence of that language, it can determine whether or not the sentence is true of the accompanying picture.

There are a number of attractive features of this general task. It is strictly behavioral and theory-free: nothing has been said about the theory or methodology that should be employed in the task. The problem is closed in the sense that one cannot appeal to some forthcoming result in a related domain that will complete the story—the system has to do the whole job. There is no stipulation of how much should be built into the system and how much learned. And the requirement that the same system should work for equivalent fragments of any natural language rules out *ad hoc* solutions.

As it turns out, a number of other groups were looking at similar tasks and there is now a modest body of work (Suppes *et al.* 1991; Nenov 1991; Siskind 1992) on the Miniature Language Acquisition task. Our own research has (as we expected) developed in ways that we didn't expect. This paper presents a brief reprise of course of the effort and a description of its current state in late 1993.

From the outset it was clear that we would not be able to attack the fully integrated L_0 task at once. The first round of effort broke it down into three projects, each of which begged a major question. The first project ignored the learning issue and concentrated on building a system that could be reprogrammed readily to solve the L_0 problem for an arbitrary natural language. There were a number of important lessons learned, primarily in the subtle semantics of spatial language, but it took well under a year to build up a hand-adaptable system (Feldman *et al.* 1990b; Weber & Stolcke 1990). Our wizards, Andreas Stolcke and Susan Weber, got to the point where it took only an hour or so to realize L_0 for a new language, given a sophisticated informant. This test-bed project was abandoned, but aspects of it are still appearing in current work.

The other two pieces of the original partition have taken a bit longer. These divided the learning aspects of L_0 into a single-word concept learning task (suppressing syntax) and a syntax learning task that assumes or ignores semantics. These turned into the doctoral theses of Terry Regier and Andreas Stolcke, respectively. We believe that the two efforts could be combined to solve the original L_0 task and we might try this as a student project, but it now seems more important to extend the task and solutions. This is best understood in the context of what has been done already. Although the concept learning system was completed earlier it is convenient to first describe the state of our progress on grammar learning.

2 Grammar Learning in L_0

Grammar induction in the context of L_0 has a two-fold function. On one hand it induces a probabilistic, syntactic model of the language, using only the provided positive samples as input. Secondly, it augments the induced syntactic model with rules for semantic features, including those representing the provided 'meaning' of the sample sentences. The resulting combined model is a probabilistic mapping that assigns both syntactic structure and meaning (within the limits of L_0 semantics) to a sentence.

Note that the way the L_0 task is posed, the syntactic aspect of learning is entirely ancillary: in principle the learning agent could construct and represent the correspondence between surface forms and semantics without any notion of syntax. However, it is generally accepted that syntactic structures have an independent motivation and are useful in both describing and processing language, although linguistic theories disagree as to the degree to which syntax should be considered 'autonomous.'

2.1 Induction of context-free grammars

Our approach to syntactic learning can be broadly described as learning of *probabilistic context-free grammars* using a *Bayesian inference* framework and *model merging* as the principal induction operator. We will explain this approach by elaborating each of these terms in turn.

Probabilistic grammars. In this framework, a grammar G is a probability distribution over all possible sequences of words, i.e., it assigns a number $P(x|G)$ between 0 and 1 to a string x, such that the probabilities of all strings sum to 1. Such an approach has several advantages over a simple all-or-nothing concept of grammar. For example, it allows objective comparisons of candidate grammars vis-a-vis a sample corpus or a given target grammar (by evaluating the sample probabilities or the relative entropies, respectively). A particularly appealing aspect of the probabilistic approach is that is provides an intuitive and formally simple solution for an old dilemma: how to learn a grammar (or more generally, any binary classification) in the absence of negative evidence, i.e., sentences labeled as non-grammatical. Under the assumption that the 'true grammar' is also probabilistic and generates the sample data according to the distribution it defines, it becomes possible in principle to identify the correct grammar (or an equivalent one) given sufficiently many data, by using the criterion above.

Context-free grammars. The particular variant of syntactic probabilistic grammars used in L_0 is the *stochastic context-free grammar* (SCFG). This is an extension of standard CFGs in which each production $X \rightarrow \alpha$ has an associated probability $P(X \rightarrow \alpha)$, the conditional probability of choosing the right-hand side α when nonterminal X is to be expanded. The context-free nature of non-probabilistic CFGs is extended by the stipulation that a nonterminal expansion is conditionally independent of the surrounding context, given X itself. This assumption is almost certainly not true of natural language, but for the fragments envisaged by L_0 it seemed to be a reasonably close model.

Bayesian inference. The probabilistic induction framework is augmented with a *prior distribution* $P(G)$ over the set of grammars. As before, each grammar G also defines a *likelihood* $P(X|G)$ relative to a set of samples X. Using Bayes Law, this allows us to obtain the *posterior probability*

$$P(G|X) = \frac{P(G)P(X|G)}{P(X)}$$

The Bayesian framework thus provides a normative approach to the inference of grammars from finite data. (In the limit of infinite data the Bayesian approach degenerates to the simple comparison of likelihoods $P(X|G)$, since that term will come to dominate the prior $P(G)$ in the equation above.)

Bayesian inference also incorporates (a version of) Occam's Razor, or a default preference for simpler over more complex models. The Bayesian version of Occam's Razor comes in two variants: for the discrete parts of grammar (i.e., its rule structure), the total prior probability mass has to be allocated over all possible grammar structures. Since the number of structural choices grows with the grammar size (measured, e.g., by the number of nonterminals, productions, etc.) each of these larger grammars will receive less prior probability than a smaller one. This is just a probabilistic reformulation of our intuition that succinct grammars are to be preferred over profuse ones, other things being equal. The second variant of Occam's Razor, know in the literature as 'Occam factors' (Gull 1988), concerns the comparison of models with continuous parameters (e.g., rule probabilities). Here, models with fewer parameters are inherently preferred by the Bayesian approach since the prior probability over the continuous parameter space has to be spread 'thinner' if that space has more dimensions. Again, this is just a mathematical reflection of our intuitive preference for models that leave less free open choices, provided they explain the data.

Within this broad characterization it is now possible in principle to construct prior distributions for grammars that reflect very specific theoretical biases. For example, we could construct a prior that assigns zero or very low probability to all context-free grammars that do not conform to the stipulations of X-bar syntax. This is known among Bayesian as an *informative prior*. On the other hand we have *uninformative priors* that simply spread the prior probability evenly in a more or less neutral fashion (clearly prior informativeness is a fuzzy concept).

We have generally taken the latter approach, since one of the interesting theoretical questions at this point is to what extent linguistic knowledge can be inferred with minimal built-in bias. Since the L_0 task is clearly a rather simple linguistic domain we would expect that a 'knowledge-free' approach would be successful here if not elsewhere.

One convenient way to come up with prior distributions on grammars is to devise ways to encode them. For example, we could encode each nonterminal in an SCFG using a certain number of bits (dependent on the total number of nonterminals), and use those codes to compile a list of all rules. The length of the entire encoding is now a precise version of what we can mean by the *size* $\ell(G)$ of a grammar G. Each coding scheme now defines a prior distribution on grammars by virtue of a standard dualism between encodings and probabilities: a grammar receives a probability proportional to $e^{-\ell(G)}$. A grammar become less probable by a constant factor > 1 for each bit more that it takes to encode it. Conversely, each prior $P(G)$ corresponds to a coding in which G is conveyed using $-\log P(G)$ bits. The Bayesian approach thereby becomes equivalent to the principle of induction by *minimum*

description length (MDL) (Rissanen 1983; Wallace & Freeman 1987). (In this framework, informative priors, or linguistic theories, can be seen as very specialized codes that provide short descriptions for all theory-conforming grammars, leaving the long descriptions for the non-conforming ones.)

Model merging. Given the Bayesian framework, it is in principle possible to find the grammar G that maximizes the posterior probability $P(G|X)$ by examining all possible grammars. Horning (1969) has shown that there is an effective procedure for doing this provided that we can enumerate all possible grammars in order of decreasing prior probability. However, such an enumeration would never be practical for any grammatical domain of even moderate size.

An alternative is to use clever heuristics that guide the search through the grammatical model space in an efficient (but not necessarily optimal) fashion. Intuitively, we would like this search to be *data-driven*, i.e., informed by the observed data, and not just blindly generating and testing hypotheses. Omohundro (1992) has recently made an argument for *model merging* as an efficient solution to this problem, which can be applied in a variety of domains and which has some degree of cognitive plausibility. The model merging approach has since been applied successfully to Bayesian inference in the domain of finite-state probabilistic grammars (Stolcke & Omohundro 1993), and our current approach for L_0 is a direct descendant of that work.

The two basic ideas in model merging are:

Data incorporation Initially, with little or sparse data, build a model that directly mirrors the data. This corresponds to a 'case-based' approach, were the best one can do is record past experiences hoping that they will be repeated in similar or identical form in the future.

Applied to the domain of language, this means that in the absence of an applicable model (or if the grammar doesn't match the new data), the learner simply adds new words, sentences, etc. to a list of exceptions yet to be explained.

Merging Whenever possible and warranted by a 'goodness' criterion, *merge*, or combine, models or submodels. The merging operation, which is dependent on the domain, usually represents a *generalization* step.

This step will be illustrated below using the L_0 domain as an example.

The incorporation step in the case of grammars is usually straightforward and involves the creation of new, specialized rules to account for the new data. For example, if our new sample is

a circle is below a triangle

and not yet accounted for by the grammar, we add a top-level production for the sentence, as well as one lexical rule for each word:

$$\begin{aligned} S &\rightarrow A_1 A_2 A_3 A_4 A_1 A_5 \\ A_1 &\rightarrow \text{a} \end{aligned}$$

$$
\begin{aligned}
A_2 &\rightarrow \text{circle} \\
A_3 &\rightarrow \text{is} \\
A_4 &\rightarrow \text{below} \\
A_5 &\rightarrow \text{triangle}
\end{aligned}
$$

The merging operation for context-free grammars comes in two variants. A *paradigmatic merging* operation involves coalescing two nonterminals. Suppose we add to the above rules one for the sentence *a circle is below a circle*:

$$S \rightarrow A_1 A_2 A_3 A_4 A_1 A_2$$

A (paradigmatic) merging step on the nonterminals A_2 and A_5 would involve replacing all occurrence of these two nonterminals with a new symbol, say, A_{25}.

$$
\begin{aligned}
S &\rightarrow A_1 A_{25} A_3 A_4 A_1 A_{25} \\
A_{25} &\rightarrow \text{circle} \\
&\rightarrow \text{triangle}
\end{aligned}
$$

As a result of merging, entire productions may be coalesced as they have become identical.

What causes 'good' merges (as the one above) to take preference over 'bad' merges (such as A_1 with A_2)? This is where the probabilistic approach becomes crucial. All productions have probabilities reflecting the relative frequencies with which each was observed in (i.e., accounted for) the data. A 'bad' merge is generally one that causes the probabilities to change in such a way as to lower the probability of the data $P(X|G)$ more than other, 'good' merges. The merge in the example above is good because it allows the two S productions to be coalesced, thereby lending high probability to the merged rule (and hence to the samples using that rule).

Alas, it is easily seen that the likelihood $P(X|G)$ can never *increase* as a result of merging; the best that can happen is that is stays unchanged. However, the Bayesian posterior probability of the grammar $P(G|X)$ does increase as a result of merging if the likelihood remains the same, since the prior will prefer a grammar with fewer nonterminals (and possibly fewer productions as well). More generally, the prior can compensate for (small) drops in the likelihood. Intuitively, this means that the data justifies a generalization beyond the observed data, due to the simplification that would bring to the grammar.

To infer general context-free grammars a second operation, *syntagmatic merging* or *chunking* is required. This involves replacing nonterminals that occur adjacently in the right-hand sides of productions by a single, new nonterminal. For example, starting with the rule

$$S \rightarrow A_1 A_{25} A_3 A_4 A_1 A_{25}$$

we are tempted to chunk the sequence $A_1 A_{25}$ to derive a new grammar containing

$$
\begin{aligned}
S &\rightarrow A_6 A_3 A_4 A_6 \\
A_6 &\rightarrow A_1 A_{25}
\end{aligned}
$$

The new nonterminal A_6 can now enter into more merging and chunking etc.

The combination of MDL prior and likelihood as our inference criterion will again distinguish 'good' from 'bad' chunks. For example, the above choice is good mainly because it cause productions to be shortened (rules with more samples in their support, such as the new $A_6 \rightarrow A_1 A_{25}$, are also preferred as a result of an Occam factor).

Summary The overall picture can now be described as follows. After formulating a prior for context-free grammars, our learner starts incorporating samples into a working grammar. It constantly examines possible merging and chunking operations, evaluating them according to the Bayesian posterior probability they produce. Operations that give improvements are applied in a best-first manner, until no more changes are warranted (and until new data arrives).

This simple description glosses over a number of non-trivial technical problems, mostly having to do with efficient search, locality of the evaluation metric etc., but we believe to have found adequate solutions to most of these. (For example, when evaluating the goodness of a chunking operation it is necessary to look ahead to the merging operations it enables to assess it true merit.)

Another interesting problem is how the learning process is affected by noise. Due to both its case-based and its probabilistic character the merging methodology handles 'outliers' quite well. They are simply added to the grammar as exception, and will eventually receive very low probability, unless supported by similar samples. Such low probability elements of the grammar can then be pruned, either on-line or in a postprocessing step.

Using simple non-informative MDL priors we have been successful in finding very reasonable syntactic description for L_0 fragments in a number of languages. For example, in the case of English, our procedure groups words into lexical categories for nouns, pronouns, transitive/intransitive verbs, adjectives, as well as the standard NP and PP phrase structures. Interestingly, the syntactic variety in L_0 seems insufficient to warrant an abstract VP category; the induced grammar keeps a flat phrase structure instead. For languages such as German the merging process correctly identifies lexical subcategories for the purpose of agreement and case assignment, and builds phrase structure rules reflecting these constraints (e.g., gender agreement between determiners and nouns).

Going beyond the L_0 project, we have recently started to use the same induction algorithm to infer context-free grammars from real natural language samples for use in a speech understanding system (Jurafsky *et al.* 1993). Preliminary results there show that the method seems to be working well, allowing us to automatically build grammars that are tightly tailored to specific patterns of language use (which is desirable for speech recognition purposes).

We have no illusions that our method can, by itself, derive arbitrarily sophisticated grammars for the more complex kinds of linguistic phenomena. Such a task would require considerable work both on the representations involved and the necessary learning algorithms. But we believe that the Bayesian model merging approach as shown for the L_0 task provides a rather general framework from with fundamental ideas can be applied to more demanding domains. Moreover, in many applications it is acceptable to use probabilistic context-free grammars as mere *approximations* to the 'true' underlying syntactic distributions. For such uses we believe to have found a very useful tool.

2.2 Learning simple feature-based semantics

In its current version the semantics of L_0 are learned *after* a context-free syntactic model has been inferred (using the methods described in the previous section, or otherwise). The tighter integration of the syntactic and the semantic aspects of learning is the focus of current research and will be briefly discussed towards the end of this section.

As in the case of syntax, learning is based on a probabilistic model for the semantics of a sentence. Even more than before, we will be interested in the discretized version of this model, but the probabilistic formulation enables an important inductive principle that is crucial to learning, namely, the idea of probability or likelihood maximization.

The sentence semantics in L_0 can be adequately represented as a collection of features. A sentence like

a circle touches a square

has a meaning represented as the following set of features names and values:

{ rel: TOUCH, tr: CIRCLE, lm: SQUARE }

where *rel* (relation), *tr* (trajector) and *lm* (landmark), as well as TOUCH, CIRCLE, SQUARE are supposedly language-independent constants of the conceptual language interfacing the language learner with the rest of L_0.

The fundamental, simplifying assumption in our approach is that these feature values originate from syntactic or lexical rules, and are combined according to rules of compositional, propositional semantics. For example, the meaning components TOUCH, CIRCLE and SQUARE are thought to be attached to the lexical rules generating the words *touch*, *circle*, *square* (the coinciding spellings are an artifact or our using English as a metalanguage). The syntactic rules that give rise to the complete sentence also contain specification that determine that these values will eventually appear on the observed top-level features.

The rules that propagate the feature values through the syntactic structure are extension of context-free rules, and similar to those found in many feature-based grammar theories (Shieber 1986).

$$\begin{aligned} X \;\rightarrow\; & Y\,Z \\ & X.f = Y.g \quad [p_1] \\ & X.f = Z.h \quad [p_2] \end{aligned}$$

This one determines that in a syntactic environment where nonterminal X expands to Y and Z, the feature f on X (briefly, $X.f$) takes on the value of $Y.g$ with probability p_1, or the value of $Z.h$ with probability p_2 (any number of such probabilistic assignments could be specified). In our current formulation, the dependence among features is a directed one: those belonging to the left-hand side nonterminal get to chose probabilistically among the values on the right-hand side, unlike non-probabilistic feature-based formalisms where the relation is a symmetrical constraint.

Feature values may also be specified directly, as is typically done at the level of lexical rules:

$$\begin{aligned} X \;\rightarrow\; & \text{circle} \\ & X.f = CIRCLE \quad [p] \end{aligned}$$

The collection of probabilistic feature rules now determines, for a given syntactic structure, the distribution of feature value appearing at the S node of the sentence. At the S node we have certain feature names determined by convention, but in the remaining parts of the structure features are entirely arbitrary; only their relation to each other is important for determining the sentence semantics.

Learning such a set of feature rules from samples (even granting knowledge of the syntactic structure) poses a chicken-and-egg problem: we could estimate the feature dependence probabilities if we knew the intermediate feature names and their values for each sample; or we could compute those values (probabilistically), if only we knew the rules.

A standard solution to this kind of problem is applicable in this scenario: the rule probabilities are initialized to random initial values, and are then optimized by repeatedly computing expected feature values and reestimating the rules so as to maximize those expectations. This procedure is generally known as the EM (expectation-maximization) method for estimating distributions with hidden variables (Dempster *et al.* 1977). (The hidden variables are the intermediate feature values in this case.) We are effectively guessing initial rules and then updating them incrementally so as to maximize the probabilities of the observed, top-level feature instances.

Applying this method to the L_0 semantics learning task, using syntactic structures of English or similar languages, we can indeed infer a system of features rules that is an accurate reflection of word meanings and their relation to sentence meanings. The algorithm typically converges on a set of probabilities that is equivalent to a discrete feature grammar, i.e., has probabilities that are either 1 or 0.

However, the method has a number of drawbacks. A general problem with the EM method is that is can find bad solutions to the optimization problem, corresponding to local, but not global, probability maxima. Depending on the initial probabilities, this can indeed happen, and seems to be more likely as the syntactic structures get deeper. Another problem is that the method requires us to fix the number of intermediate features in advance (although redundancy cannot hurt). In current work we are trying to overcome both of these difficulties by integrating feature estimation into the syntactic induction process. The basic idea here is that simple feature associations between words and sentence meanings are more easily found in the early stages of grammar learning while the syntactic structures are still flat. As merging and chunking are applied to the syntax, corresponding feature merging and bundling operation are performed. Importantly, the effect that syntactic changes have on the feature rules can now help in deciding between good and bad changes in the structure of the grammar. This will allow us to build a simple model of syntactico-semantic synergy during learning.

3 Lexical Spatial Semantics

3.1 The Linguistic Categorization of Space

The linguistic categorization of space is a topic that has captured the attention of linguists and other cognitive scientists for a number of years. There is good reason for this: space is an attractive semantic domain in that it is concrete and tangible, and yet still affords a good deal of subtlety in its semantic structure. In this respect it resembles the domain of color,

Mixtec *siki*
English *above*

Mixtec *šini*
English *above*

(a)

German *auf*
English *on*

German *an*
English *on*

(b)

Figure 1: Examples from Mixtec and German

---

another objectively measurable domain which has attracted work in semantics (Berlin & Kay 1969).

In addition, space (along with time) has a privileged position as a fundamental conceptual structuring device in language, a position which most other domains, color included, do not share. Lakoff (1987) points out that such fundamental concepts as those relating to time and space tend to be used in many other concepts throughout a given linguistic system, often through metaphor, and are not localized to isolated domains of experience, as most other concepts (e.g. *socialism*) are.

Space is also attractive as a domain of semantic inquiry because there is considerable cross-linguistic variation in spatial systems. The notion of cross-linguistic variation in conceptual systems and modes of thought has long been a subject of fascination for linguists (see Whorf (1956) for an early and extremely influential example). Such differences in spatial systems are sometimes quite dramatic, but more often than not they are rather subtle, particularly when one compares closely related languages. Figure 1 provides examples of non-English spatial structuring, from Mixtec (Brugman 1983), a Mexican Indian language, and from German. In Figure 1(a) we see two spatial configurations which would both be categorized as *above* in English, but which receive distinct categorizations in Mixtec, which is sensitive to the major axis orientation of objects. Similarly, in (b) we see two configurations both of which would be categorized as *on* in English, but which are treated distinctly in German, in which the orientation of supporting surfaces is semantically significant. An exhaustive cataloging of cross-linguistic variation in spatial systems is of course well beyond the scope of this article. There are however at least two fairly large-scale studies of cross-linguistic variation in spatial terms underway, one by Goldberg & Slobin (1992), and the other by Eric Pederson and Melissa Bowerman (personal communication).

Despite all this variation, there must also exist commonalities across languages by virtue

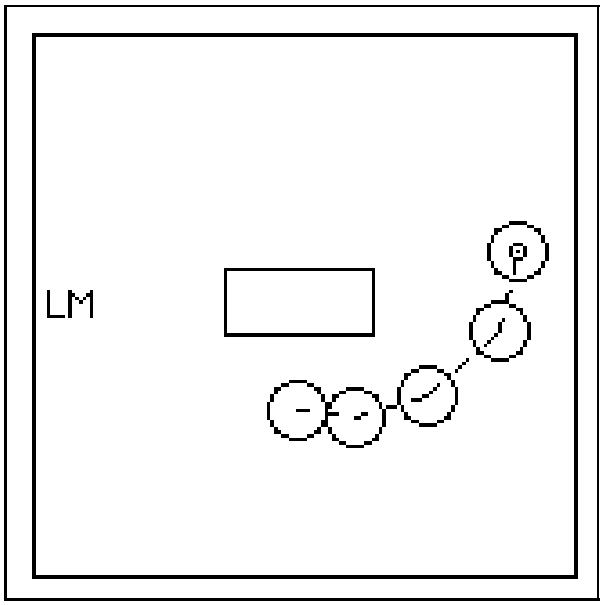

Figure 2: A movie: Russian *iz-pod*

---

of the fact that all linguistic spatial systems are based on human experience of space, which is in turn constrained by the nature of the human perceptual system and the nature of the world around us. The point of interest here is that the varying spatial systems all derive from the same neural mechanisms and the same experiences with objects, gravity, and the like.

This leads us to ask the following question:

> *What sort of system could adapt itself to so many different structurings of space, as the human capacity for categorizing space can?*

We present a connectionist model which learns the perceptually grounded semantics of lexemes for spatial events and relations from a range of languages. Thus, it provides at least a preliminary model of the human capacity for lexicalizing spatial events and relations, and a preliminary answer to the above question.

3.2 The Model

Imagine a set of movies of simple 2-dimensional objects moving relative to one another, such that each movie has been correctly labeled as a positive instance of some natural language spatial term. The movies may be of arbitrary length. For example, Figure 2 presents such a movie, a positive example of the Russian preposition *iz-pod*, which has no single-word English counterpart, but translates to "out from underneath". The connectionist model presented here takes a set of such movies, each labeled as a positive example of some spatial term from some language, and learns the association between words and the events or relations which they describe. Once such a system has successfully accomplished this task,

it should be able to determine which of the spatial terms it has learned would be appropriate for describing a movie not previously seen. The system's task as a whole can be viewed as:

> *Learning how to perceive simple spatial relations, both static and dynamic, so as to name them as a speaker of a particular language would.*

Clearly, if this system is to be taken as a model of the human capacity for lexicalizing spatial relations and events, it must be able to perform this learning task for spatial terms from any language. Furthermore, it should be able to learn without the benefit of explicit negative instances, as children appear to acquire language under those conditions. While it is unlikely that the model as currently constructed can learn an arbitrary human linguistic spatial system, the model does learn systems of spatial terms from a range of languages with disparate categorizations of space. And it learns from positive evidence only. Thus it begins to answer the question posed above: it provides an indication of what sort of system could adapt itself to the various structurings of space found in the world's languages.

The key to the model's design is its incorporation of a number of structural devices which are motivated by neurobiological and psychophysical evidence concerning the human visual system. The model contains motivated structures of three sorts: (a) orientation-sensitive cells, (b) cells with circularly symmetric center-surround receptive fields, and (c) a tripartite representation of motion trajectories which consists of separate subrepresentations for the starting point, ending point, and path taken.

There is evidence for each of these structures in the human visual system. The critical point is that the incorporation of these structures into the model is in keeping with the spirit of inquiry into issues of semantic universality and relativity that motivates the work as a whole. After all, the central tension is between cross-linguistic variation on the one hand, and the knowledge that languages must have some semantic structure in common simply because all humans have essentially the same perceptual system. The model presented here fits in as an element of this line of inquiry, as it is a linguistic learning system whose architecture is inspired by neurobiological and psychophysical insights into the workings of the human visual system.

Figure 3 presents the model's overall architecture, with the non-linguistic structures we have been discussing highlighted. As shown in the figure, the model is configured to learn a set of English spatial terms. Once a movie has been shown to the trained network, the nodes corresponding to those lexemes which describe the event shown should be fully activated. The model is trained using the connectionist error back-propagation training algorithm (Rumelhart *et al.* 1986), while its architecture is informed by the design philosophy of *adaptive structured connectionism* (Regier & Feldman 1993). The fundamental idea behind this philosophy is the combination of elaborate innate structure, such as that found in actual neural structures in the brain, with the remarkable adaptability that characterizes their operation.[1] The model illustrated in Figure 3 is an example of this: the neurobiologically and psychophysically motivated structures mentioned above are embedded in a

[1] The general idea of structuring connectionist learning networks is not new; see for example (LeCun 1989; Keeler *et al.* 1991; Mozer *et al.* 1991; Guyon *et al.* 1991), among others. What distinguishes adaptive structured connectionism is its emphasis on the use of highly domain-specific structure, such as that focused upon here, rather than very general techniques such as weight-sharing across temporal or spatial locations.

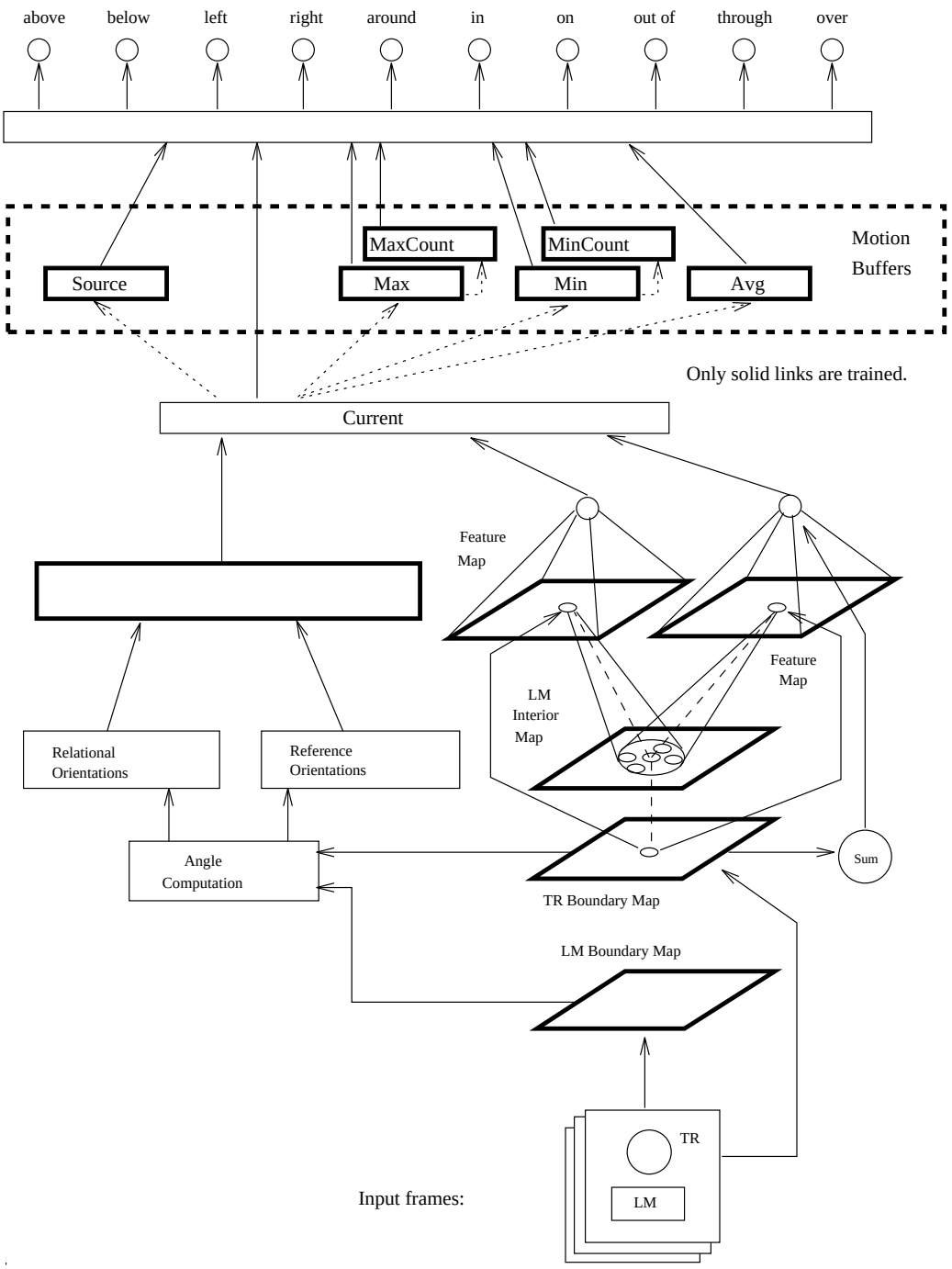

Figure 3: The model's architecture, with non-linguistic structures highlighted.

trainable network whose parameters are adjusted during training to come to reflect the spatial structuring of a particular language.

Although we shall not be able to describe this network in detail, a few words regarding its operation and ramifications are in order. The model has learned the perceptual grounding for spatial lexemes from English, Mixtec, German, Bengali, and Russian, and preliminary testing has yielded encouraging results for Japanese as well. For all languages, generalization to test sets is excellent. The model is able to learn without the benefit of explicit negative evidence, and as a result of its structure and operation, gives rise to a number of linguistic predictions concerning which semantic features one may or may not expect to find in spatial lexemes in the world's languages. We refer the interested reader to (Regier 1992; Regier 1993) for further details.

4 Current and Planned Efforts

The problem of learning and representing spatial relations terms in the world's language is extraordinarily difficult, and Regier's model is only a beginning. It is limited to 2-dimensional representations, and to artificial scenes with geometrical figures. In addition, it lacks force dynamics, deixis, object grouping, imposed framing, polysemy, and much much more. Like any other fully explicit model, it makes one realize just how difficult the job is.

Regier's model is, of course, limited to spatial relations concepts and does not include any of the metaphorical apparatus necessary to characterize abstract concepts. At present there is no serious model of any significant portion of the system of metaphorical mappings needed to characterize abstract concepts.

Most important, Regier's model does not characterize spatial inferences. To do so, it would need to include some ability to do binding. Such capabilities are available in connectionist models, but until some form of binding is incorporated into some version of Regier's model, this gap remains. There are also a number of other problems with inference in Regier's model discussed below. Both of the current semantic projects focus on extending the representation from layers of units to more active recurrent connectionist networks. It is not clear how far we can go with the existing models, but we are convinced that some version of the schema story is needed for further progress.

4.1 Extension to Hand Domain

4.1.1 Problem Definition

A new project is underway which extends the domain of the original L_0 problem definition. David Bailey's thesis work involves the learning of action-oriented terms which apply to activities performed with the hand, such as "get", "push", "open", and perhaps "stop" or "woops". It is expected that the semantics of these terms can be expressed (and learned) in terms of suitably-defined, relatively high-level action "schemas"—essentially compiled plans—which we imagine children to have already learned from experience when they come to the language learning task.

The goal, then, is to be able to build a system, containing a set of schemas for common hand actions, which can learn the semantics of hand-action terms for any language, from

labelled examples of schema executions in simulated world configurations. The learned knowledge should be demonstrable by the system's carrying out simulated verbal commands involving the learned terms in novel world configurations. Furthermore, we would like to build a system to learn these schemas from goal-driven experimentation.

4.1.2 Schemas

Activity sequences and plans seem to be characterized by "key events" which cause a transition from one action to another. Furthermore, it appears that atomic primitive actions suffice at this level of representation. For example, a GET-OBJECT schema may work as follows:

> "Concurrently move hand in direction of object and open fist, until contact occurs. Then close fist until all five fingers are exerting pressure p. Then move hand up and toward self."

Thus, we envision representing schemas as some type of finite-state device. Several extensions to FSMs seem necessary. First, since multiple primitive actions can take place simultaneously, we need a simple mechanism for concurrency. Second, we would like our agent's set of schemas to be hierarchical to maximize reuse of simple task strategies when implementing complex tasks, so we plan to utilize the standard augmented transition network (ATN) strategy of allowing actions on transitions to invoke entire (sub-)schemas. Third, we plan to parameterize our schemas along various dimensions, such as object size, speed of movements, etc. The parameters are used by the schema to compute parameters of its primitive actions and sub-schemas.

In many cases, action terms will be expressible as schema invocations with particular parameter settings. As an example, consider the following extract from Webster's Online:

> "TAKE, SEIZE, GRASP, CLUTCH, SNATCH, GRAB mean to get hold of by or as if by catching up with the hand. TAKE is a general term applicable to any manner of getting something into one's possession or control; SEIZE implies a sudden and forcible movement in getting hold of something tangible or an apprehending of something fleeting or elusive when intangible; GRASP stresses a laying hold so as to have firmly in possession; CLUTCH suggests avidity or anxiety in seizing or grasping and may imply less success in holding; SNATCH suggests more suddenness or quickness but less force than SEIZE; GRAB implies more roughness or rudeness than SNATCH."

If our GET-OBJECT schema were parameterized as GET-OBJECT<speed, force, security, precision>, we might translate the above set of terms as follows:

	speed	force	security	precision
take	any	any	any	any
seize	HIGH	HIGH	any	any
snatch	HIGH	LO/MED	any	any
grasp	any	any	HIGH	any
grip	any	any	HIGH	any
grab	any	any	any	LOW
clutch	any	HIGH	LOW	any

Note we are not able to make all distinctions—more on this below.

4.1.3 Motivation

This extension to L_0 is interesting for several reasons:

1. A traditional contrast in AI exists between "procedural" and "declarative" forms of knowledge. The procedural form facilitates control, while the declarative form supports flexible reasoning. A challenge presented by this task is that our schemas must support both uses; they must be able to drive (simulated) behavior, but must also be amenable to learning language which maps to various aspects of the activities they encode.

2. The requirement that schemas must themselves be learnable presents another pair of opposing constraints on their representation: for lexical semantics we prefer a rich, detailed representation; but for schema learning, we prefer the simplest possible representation. A goal is to find a midpoint which is tractable for learning but sufficient for language.

3. Another area to explore is how the two learning tasks interact. The order of lexical acquisition may depend on and be partially explained by ongoing schema development. Moreover, the language learning task may be able to capitalize on starting out with a smaller schema set. More interestingly, language learning itself may cause re-organization of the schema knowledge, leading to Whorfian effects that would be interesting to model or predict. For example, when an action term becomes associated with two distinct schemas (say, "pull", PULL-BY-ROPE, and PULL-BY-HANDLE), there may follow a consolidation into a single parameterized schema, facilitating the sharing of strategies by the two (say, PULL-BY-ANY-GRABBABLE-PART).

4. Lastly, this project connects with Srini Narayanan's thesis work (described below) involving the mapping of narratives in the "abstract" domain of politics and economics into activities of the body, so that abstract inferences and predictions may be made via this mapping to bodily knowledge. We are hoping that similar schema representations will suffice for both projects.

4.1.4 Status

At this point work is proceeding along four fronts:

1. We are in a continual process of collecting psychological and linguistic data. Questions that the psychological literature may shed light upon include, "What tasks do children first learn to perform with their hands?", and "How to children's problem solving strategies evolve in the first two years?". From the linguistic literature (and a bit of our own data-gathering) we hope to answer the questions, "Which action-terms are learned first?" and "What kinds of distinctions are made by action-terms in different languages?". This last question is particularly important—while we know that we cannot represent all nuances of our action terms with our schemas and their numerical and geometric parameters (in our example, we cannot capture the "avidity" of clutch nor the "rudeness" of grab), it is important to find out how close we can come with such a substrate. Tomasello (1991) provides the kind of developmental data we need.

2. A key feature of earlier L_0 work was its use of an abstract world of two dimensional line-drawings of simple geometric shapes, to sidestep hard problems in vision so that we could focus our mental and computational resources on higher-level issues, while maintaining a plausible connection to the real-world problem faced by humans. In switching to the hand-action domain, we must likewise avoid simulating detailed robotics, while still capturing enough of human hand activity so that psycho-linguistic data is pertinent. Taking our cue from simplifications made in the robotics world (Venkataraman & Iberall 1990), we envision simulating a five-fingered hand with one degree of freedom on the various finger joints. We will abstract away the arm, using a primitive to move the hand to any position and orientation. Each simulated finger will have a real-valued force sensor, and proprioceptive sensors will provide joint positions, velocities, and forces. For actions, we'll assume the hand can carry out simple actions such as "move hand forward until contact" as a primitive (the motivation here is that we will encapsulate those activities which we imagine are controlled by low-level, unconscious neural circuitry in humans). The outside world will consist of objects—described by constructive solid geometry plus a handful of other properties such as surface friction and mass—placed on a tabletop in a gravitic environment. We are looking into strategies used in the qualitative physics and animation communities in order to develop our simulator.

3. We will devise learning techniques that can support the synthesis of schemas from either free experimentation in the world-simulator, or from training examples of successful goal-achievements (being led by hand, so to speak). One possibility is a reinforcement learning approach. Here, the central problems are (1) including agent state, and (2) input generalization. Whitehead (1992) presents an interesting technique for the latter problem based on training an active perception mechanism along with learning the state-action mapping. The ATN-like aspect of schemas suggests a very different approach, namely, to reuse portions of Stolcke's thesis work on grammar induction. A difficulty here, though, is that our task requires learning *conditions* on rules, not probabilities. Lastly, we will consider Goddard's (1992) connectionist representation of "scenarios" for human walking, and how they might be learned.

4. Perhaps the central learning task in this project is the mapping from action terms to schemas. It is hoped that with a proper schema design, this learning problem will be straightforward. In the simplest case, action terms which refer to achievement of a goal may simply map to any invocation of an existing schema designed to achieve that goal. Other terms may generalize to multiple schemas. In other cases, action terms impose certain constraints on achievement of a goal which may be expressible as constraints on the parameters of the schema (as with the "take"-words above). More complex terms may refer to continuous ("slide"), repetitive ("wave"), or patterned ("flick") behaviors occurring within any schema. Yet other terms may map to generalized properties of control flow of schemas, such as "stop" or "woops".

4.2 Metaphor

4.2.1 Introduction

This project attempts to build a computational model of the semantic structure of causal narratives. The input is taken from newspaper stories in the domains of Politics and Economics, and the proposed system should represent the semantics in a way that can generate commonsense inferences about the stories.

Two observations motivate this project. First, causal narratives from abstract domains acquire a significant portion of their semantics from metaphoric mappings to specific schemas in the concrete spatial and temporal domains. We postulate therefore, that an adequate representation of the concrete domain is an essential component of modeling the semantics of these narratives. Recent work in Cognitive Semantics provides good evidence for this view. In particular, there are proposals for schema-based semantics, as in the case of *Image Schemas* (Lakoff 1987). The work in Cognitive Semantics, however, lacks a computational model for such theories.

Second, we note that the deep semantics of the causal narratives are *dynamic* and arise from a *continuous interaction* between input and memory. This enables the high degree of context sensitivity required since changing input context can dramatically affect the correlation between input and memory and thereby the set of possible expectations, goals, and inferences. Any representation of the semantics should therefore be of a fine granularity, be flexible and adaptive in the way context affects interpretation, and be able to support concurrent activities in the memory storage, retrieval, and indexing process.

Our observations suggest the use of *process semantics* (Nielsen *et al.* 1981) to characterize the deep semantics of causal narratives. Of special interest to us is the possibility of modeling such processes in structured connectionist models (Feldman 1989).

This work attempts to develop computational models based on a synthesis from process theory, insights from structured connectionist systems and from linguistic research on Cognitive Semantics to model the commonsense understanding of causal narratives.

We first illustrate these observations by describing their usage in interpreting a recent newspaper story. We then describe the status of this effort and the problems remaining. Finally the scope of the project is summarized.

4.2.2 Example

Consider the following paragraph that appeared in the New York Times on Sunday, August 1, 1993.

> Britain was deep in recession while Germany was booming three years ago. France kept booming long after Germany sunk into recession. Now France is plunging deeper while Germany hopes it has bottomed out. Britain, which was able to cut interest rates after leaving the exchange rate mechanism, has been able to stimulate its economy and has started to emerge from recession.

In interpreting the narrative above, there are several aspects to be noted.

- Germany, France, and Britain are all conceptualized as persons. This is a fairly common occurrence and is based on a conventional metaphoric mapping *Country Is Mapped As Person*.
- Economic recession is treated as a feature in a terrain, namely a *hole* that you can sink into, plunge into, be deep in, emerge from, and a *hole* which can bottom out. The mapping from an abstract economic state (such as Recession) to a concrete spatial location/region is an instance of the metaphoric mapping *Abstract State Is Mapped As Spatial Region*. This particular metaphor belongs to a set of inter-related metaphors called The Event Structure Metaphor (Lakoff 1992).
- There is an implied inference about the *negative* aspects of sinking into holes, and the *positive* aspects of booming or emerging out of them. This kind of directional preference is an instance of a class of *orientational* metaphors. Notice that this preference generates an *implicit goal* for any agent, that constrains the possible interpretations of the input text.
- From the fact that the countries are conceptualized as persons, we note the specific causal events that are used in the article, such as sink into, emerge from, plunge into, stimulate, leave, etc. These are specific *agent motions in space* whose semantics is based on the *manner* and of motion, the implied or explicit *path* (such as *into*), the *resultant state* (such as increased energy in stimulate). In general, there may be other relevant features (such as *aspect*), which contribute to the meaning of the narrative.

4.2.3 Modeling Approach

Described below are the inputs and outputs of the proposed model followed by a summary of the current status of the effort.

Input/Output

- **Input**.

  We assume as input surface parsed syntactic and semantic elements from newspaper stories in the domains of politics and economics. Surface syntactic elements include

the constructs "verb", "adposition", "subordinate clause", VP, NP, PP and "satellites" (Talmy 1985). Surface semantic elements include "Figure", "Action","Action Modifier", "Figure Modifier", Ground Modifier". The system will include a lexicon of the relevant action terms and the associated surface semantic role constraints. Examples of action terms include carry, lead, plunge, push, touch, back, cripple, fall, and stumble.

- **Output**.

  The output of the proposed system will be the set of bindings that result from interpreting the abstract narrative as a series of mappings to schemas in the spatial and temporal domain. The mappings make use of metaphors (such as the Event Structure Metaphor) as well as the specific process characteristics resulting from the execution of the concrete domain schemas.

Current Status Our approach to the modeling of the metaphoric understanding is proceeding along the following dimensions.

- Identification and representation of the set of conventional metaphors that are relevant in the understanding of social causality.
- Representation of the concrete spatial domain. Apart from the general requirement of being able to model states (with real variables) and transitions (including hierarchical transitions) we have identified some specific requirements that pertain to modeling concurrent activation and execution of multiple schemas. These specific features are outlined below.

  1. Should be able to distinguish *enabling* of actions from their *execution*.
  2. Should support specific process primitives such as *temporal order*, *enabling/disabling*, *conflict*, *choice/nondeterminism*, and *covarying actions*.
  3. Should be implemented as a connectionist style of computation (ie. should be based on finite state machines, and should use a generalized spreading activation mechanism for inference).

  The requirements outlined above suggest a net based model of process semantics. The requirements of learnability and connectionist implementation require us to define special modifications to standard concurrency models (such as Petri Nets and Event Logics (Nielsen *et al.* 1981)).

- Mapping the abstract social domain structures into the relevant concrete spatial and experiential domain structures. This includes the selection, initialization and execution of specific concrete domain schemas. To this end, we have divided the domain schemas into the following categories:

  – **The World Model**. This is the overall model of the characteristics of the spatial domain. The world simulation comprises of a model of space called a *spatial map* (that includes relevant terrain features such as holes, bumps, valleys, etc.). The

proposed world model also includes an agent that moves through the spatial map affected by various agent initiated as well as external forces (gravity, wind, and friction).

- **The Body Model**. The body model comprises of a schema that models various body parts, spatial relations between these parts, and various postural constraints (based on relevant joint location, force and motion constraints). We assume that a *CSG* representation of the body along with a discrete event characterization of the dynamics is sufficient for our purpose.
- **The Social Cognition Model**. The social cognition model pertains to specific family and social relationship schemas. The various aspects of this model are currently being developed.

4.2.4 Summary

This project proposes to investigate some new ideas for modeling the understanding of narratives such as newspaper stories involving political or economic causation. The central novel ideas proposed here are

1. A computational model of narrative understanding by metaphoric mapping from abstract domains, such as politics and economics to the concrete spatial domain.

2. The translation of the metaphoric interpretation process into multiple, executing models or schemas. This allows for a single system to be query answering, generative and capable of counterfactual reasoning.

3. The grounding of the deep semantics of the abstract causal terms in body-based active models.

5 Synthesis and Conclusion

The theme of this issue is that much can be gained from considering vision and natural language processing problems together. The L_0 project has extended this idea to encompass learning and, more recently, motor control and metaphorical inference. We believe that we have already shown that this broad approach can yield valuable insights and, somewhat surprisingly, practical applications as well. We began by looking hard at the task of first language acquisition in the domain of simple 2D geometric scenes. This task divided cleanly into a concept learning problem and one of inducing grammars from sample sentences. Notice that both sub-problems span issues in learning, language, and spatial representation. Despite considerable initial success, we were unable to extend our paradigm to encompass simple spatial inferences that should be part of knowing the meaning of a term. Current work aims at developing representations capable of supporting learning and inference and also usable as a base for metaphoric mappings from other domains.

Another interesting aspect of the work to date is the range of methodologies used. Although we were originally inclined to attack the L_0 task with connectionist methods on all fronts, we quickly conceded to ourselves that that was not realistic, at least not given the current state of the field. The initial L_0 testbed (Feldman *et al.* 1990b) relied on traditional

symbolic AI methods for purposes of rapid prototyping. The grammar learning project started out connectionist (Stolcke 1990), but eventually adopted a probabilistic framework which was thought to provide more versatile representations while still capturing the 'soft computation' aspects of the problem. The spatial semantics learning project has remained the most purely connectionists. The current projects are informed by, but not in any way committed to connectionist solutions.[2]

The entire L_0 project should be viewed as part of a larger movement to develop a grounded and unified cognitive science. We start with the assumption that Cognitive Science, like other sciences, should be grounded and that base for the grounding must be the physical and biological sciences. The ultimate determiner of our minds is the situation of our existence, co-existence, and evolution on the planet. Most importantly, we claim that taking this stance seriously now can lead to a unified and more productive Cognitive Science.

The paradigm example of what we have in mind is the work on color-terms in the world's languages and its basis in physiology and psychophysics. Decades of work by people from a wide range of disciplines has produced an understanding of this (admittedly very narrow) aspect of cognition that seems as solid and incontestable as any in the sciences (Berlin & Kay 1969). Our claim is that this paradigm can be extended, albeit with enormous effort, to yield a Cognitive Science that will establish a continually growing body of established scientific truth.

The central idea is to take very seriously the notion that all of our concepts and cognitive processes have their roots in the nature of our bodies and brains. The color story provides a clear, albeit deceptively simple, version of the fruitfulness of this approach. There are a variety of existing projects that are providing pieces of similar stories in other basic domains. Connectionist models provide a scientific language for linking concrete concepts to the body. The path to treating more abstract domains is being explored in studies of metaphor and other mappings from abstract to concrete representations.

This is not to suggest that everyone should immediately start working at the foundational level. No field has ever developed this way. What is crucial is to recognize that there is an incontestable core of science on which all cognitive theories must eventually be based. It is always relevant to ask the current state of the grounding and it is necessary to be skeptical when no answer is forthcoming.

References

Berlin, Brent, & Paul Kay, 1969. Basic color terms: Their universality and evolution. University of California Press, Berkeley.

Brugman, Claudia, 1983. The use of body-part terms as locatives in chalcatongo mixtec. in Report No. 4 of the Survey of California and other Indian Languages, pp. 235-90. University of California, Berkeley.

[2]We might add that the methodological development of the L_0 project is somewhat symptomatic of connectionist research as a whole during recent years.

Dempster, A. P., N. M. Laird, & D. B. Rubin. 1977. Maximum likelihood from incomplete data via the *EM* algorithm. *Journal of the Royal Statistical Society, Series B* 34.1–38.

Feldman, Jerome A. 1989. Neural representation of conceptual knowledge. In *Neural connections, mental computation*, ed. by Lynn Nadel *et al.*, 68–103. Cambridge, Mass.: MIT Press.

——, George Lakoff, Andreas Stolcke, & Susan Hollbach Weber. 1990a. Miniature language acquisition: A touchstone for cognitive science. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*, 686–693, MIT, Cambridge, Mass.

——, Susan Hollbach Weber, & Andreas Stolcke. 1990b. A testbed for the miniature language L_0. In *Proceedings of the 5th Rocky Mountain Conference on Artificial Intelligence*, 25–30, New Mexico State University, Las Cruzes, N.M.

Goddard, Nigel, 1992. *The perception of articulated motion: Recognizing moving light displays*. University of Rochester dissertation.

Goldberg, Adele, & Dan Slobin, 1992. A cross linguistic study of spatial terms. Technical report draft, International Computer Science Institute.

Gull, S. F. 1988. Bayesian inductive inference and maximum entropy. In *Maximum entropy and Bayesian methods in science and engineering, volume 1: Foundations*, ed. by G. J. Erickson & C. R. Smith, 53–74. Dordrecht: Kluwer.

Guyon, I., P. Albrecht, Y. LeCun, J. Denker, & W. Hubbard. 1991. Design of a neural network character recognizer for a touch terminal. *Pattern Recognition* 24.105–119.

Horning, James Jay. 1969. A study of grammatical inference. Technical Report CS 139, Computer Science Department, Stanford University, Stanford, Ca.

Jurafsky, Daniel, Chuck Wooters, Gary Tajchman, Jonathan Segal, Andreas Stolcke, & Nelson Morgan. 1993. The Berkeley Restaurant Project. Technical report, International Computer Science Institute, Berkeley, CA. To appear.

Keeler, James, David Rumelhart, & Wee-Kheng Leow. 1991. Integrated segmentation and recognition of hand-printed numerals. Technical Report ACT-NN-010-91, Microelectronics and Computer Technology Corporation.

Lakoff, George. 1987. *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago Press.

——. 1992. What is metaphor? In *Advances in connectionist and neural computational theory, Vol. 2: Analogical connections*, ed. by K. J. Holyoak & J. A. Barnden. Ablex.

LeCun, Yann. 1989. Generalization and network design strategies. Technical Report CRG-TR-89-4, Connectionist Research Group, University of Toronto.

Mozer, Michael, Richard Zemel, & Marlene Behrmann. 1991. Learning to segment images using dynamic feature binding. Technical Report CU-CS-540-91, Dept. of Computer Science, University of Colorado at Boulder.

Nenov, Valeriy I. 1991. Perceptually grounded language acquisition: A neural/procedural hybrid model. Technical Report CSD-910083, Computer Science Department, University of California, Los Angeles.

Nielsen, M., G. Plotkin, & G. Winskel. 1981. Petri nets, event structures, and domains, Part I. *Theoretical Computer Science* 13.

Omohundro, Stephen M. 1992. Best-first model merging for dynamic learning and recognition. In *Advances in Neural Information Processing Systems 4*, ed. by John E. Moody, Steve J. Hanson, & Richard P. Lippman, 958–965. San Mateo, CA: Morgan Kaufmann.

Regier, Terry, 1992. *The acquisition of lexical semantics for spatial terms: A connectionist model of perceptual categorization.* Computer Science Division, EECS Department, University of California at Berkeley dissertation. available as Technical Report TR-92-062, International Computer Science Institute, Berkeley.

——. 1993. Two predicted universals in the semantics of space. In *Proceedings of the Nineteenth Annual Meeting of the Berkeley Linguistics Society.* University of California, Berkeley.

——, & Jerome Feldman, 1993. Structured connectionist models and spatial concept learning. In J. Hendler, editor, *Handbook of Neuropsychology, Volume 9* (to appear). Elsevier.

Rissanen, Jorma. 1983. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics* 11.416–431.

Rumelhart, David E., Geoffrey E. Hinton, & Ronald J. Williams. 1986. Learning internal representations by error propagation. In *Parallel distributed processing: Explorations in the microstructure of cognition*, ed. by James L. McClelland & David E. Rumelhart, 318–362. MIT Press.

Shieber, Stuart M. 1986. *An introduction to unification-based approaches to grammar.* Number 4 in CSLI Lecture Note Series. Stanford, Ca.: Center for the Study of Language and Information.

Siskind, Jeffrey Mark, 1992. *Naive physics, event perception, lexical semantics, and language acquisition.* Cambridge, Mass.: Massachussetts Institute of Technology dissertation.

Stolcke, Andreas. 1990. Learning feature-based semantics with simple recurrent networks. Technical Report TR-90-015, International Computer Science Institute, Berkeley, CA.

——, & Stephen Omohundro. 1993. Hidden Markov model induction by Bayesian model merging. In *Advances in Neural Information Processing Systems 5*, ed. by Stephen José Hanson, Jack D. Cowan, & C. Lee Giles, 11–18. San Mateo, CA: Morgan Kaufmann.

Suppes, Patrick, Lin Liang, & Michael Böttner. 1991. Complexity issues in robotic machine learning of natural language. In *Modeling complex phenomena*, ed. by L. Lam & V. Naroditsky. New York, N.Y.: Springer Verlag.

Talmy, Len. 1985. Lexicalization patterns, semantic structures in lexical form. In *Language typology and syntactic description*, ed. by Timothy Shopen. Cambridge: Cambridge University Press.

Tomasello, Michael. 1991. *First verbs.* Cambridge University Press.

Venkataraman, S.T., & T. Iberall. 1990. *Dextrous robot hands.* Springer–Verlag.

Wallace, C. S., & P. R. Freeman. 1987. Estimation and inference by compact coding. *Journal of the Royal Statistical Society, Series B* 49.240–265.

Weber, Susan Hollbach, & Andreas Stolcke. 1990. L_0: A testbed for miniature language acquisition. Technical Report TR-90-010, International Computer Science Institute, Berkeley, CA.

Whitehead, Steven, 1992. *Reinforcement learning for the adaptive control of perception and action.* University of Rochester dissertation.

Whorf, Benjamin Lee. 1956. *Language, thought, and reality.* Cambridge, MA: MIT Press. (ed. John B. Carroll).